HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Phylogenomics and *Adaptive Evolution* in Stickleback Fishes

Bohao Fang

2020

FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

Phylogenomics and adaptive evolution in stickleback fishes

Bohao Fang

Ecological Genetics Research Unit

Organismal and Evolutionary Biology Research Programme

Faculty of Biological and Environmental Sciences

University of Helsinki

Finland

Academic Dissertation

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the Univeristy of Helsinki in Auditorium 2041, Biocenter 2, Viikinkaari 5 on December 10th at 12 noon

Supervised by:

Prof. Juha Merilä
Organismal and Evolutionary Biology Research Programme, University of Helsinki, Finland

Dr. Paolo Momigliano
Organismal and Evolutionary Biology Research Programme, University of Helsinki, Finland

Thesis advisory committee:

Prof. Päivi Onkamo
Department of Biology, University of Turku
Organismal and Evolutionary Biology Research Programme, University of Helsinki, Finland

Dr. Ari Löytynoja
Institute of Biotechnology
University of Helsinki, Finland

Reviewed by:

Prof. Michael M. Hansen
Department of Bioscience
Aarhus University, Denmark

Prof. Walter Salzburger
Zoological Institute
University of Basel, Switzerland

Examined by:

Prof. Roger Butlin
Department of Animal and Plant Sciences
University of Sheffield, UK
Department of Marine Sciences
University of Gothenburg, Sweden

Custos:

Prof. Marjo Saastamoinen
Helsinki Institute of Life Science
University of Helsinki, Finland

Cover design: Bohao Fang
Art & typesetting: Bohao Fang
ISBN 978-951-51-6764-4 (paperback)
ISBN 978-951-51-6765-1 (PDF)
http://ethesis.helsinki.fi
Unigrafia, Helsinki, 2020

2016

2020

BOHAO FANG

方博豪

于芬兰赫尔辛基

CONTENTS

LIST OF ORIGINAL ARTICLES

The thesis is based on the following chapters, which are referred to in the text by their Roman numerals:

I **Fang B**, Merilä J, Ribeiro F, Alexandre CM, Momigliano P (2018). Worldwide phylogeny of three-spined sticklebacks. *Molecular Phylogenetics and Evolution* 127, 613-625.

II **Fang B**, Merilä J, Matschiner M, Momigliano P (2020). Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Molecular Phylogenetics and Evolution* 142, 1055-7903.

III Guo B*, **Fang B***, Shikano T, Momigliano P, Wang C, Kravchenko A, Merilä J (2019) A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus *Pungitius*. *Molecular Ecology* 28, 4046-4064.

IV **Fang B***, Kemppainen P*, Momigliano P, Feng X, Merilä J (2020). On the causes of geographically heterogeneous parallel evolution in sticklebacks. *Nature Ecology and Evolution* 4, 1105-1115.

V **Fang B**, Momigliano P*, Kemppainen P*, Merilä J (2020). Population structuring limits parallel evolution. *Manuscript*.

** Equal contribution*

AUTHORS' CONTRIBUTIONS

CHAPTERS	I	II	III	IV	V
Concept	JM, PM, **BF**	PM, **BF**, JM	JM, BG, TS, PM, **BF**	PK, JM, PM, **BF**	JM, **BF**, PM, PK
Analyses	**BF**, PM	**BF**	**BF**, BG	**BF**, PK, PM	**BF**, PK, PM
Manuscript preparation	**BF**, PM, JM	**BF**, PM, JM	JM, BG, **BF**	PK, **BF**, PM, JM	JM, **BF**, PM, PK

BF: Bohao Fang JM: Juha Merilä PM: Paolo Momigliano

PM: Petri Kemppainen BG: Baocheng Guo TS: Takahito Shikano

Abstract

How predictable is evolution? There is no fully satisfactory answer to this 100-year old question yet. However, within the past two decades, much progress has been made towards unravelling various factors that influence the predictability of evolution. Much of this work has focused on the similarity of evolutionary responses in replicate populations of a given taxon that have independently colonised similar environments – a phenomenon known as parallel evolution. The fish species in the family Gasterosteidae (sticklebacks) have become popular models to study the repeatability of evolution.

This thesis focuses on evolutionary history and parallel evolution in two ecologically similar and geographically co-distributed species in the family Gasterosteidae, the three-spined stickleback (*Gasterosteus aculeatus*) and the nine-spined stickleback (*Pungitius pungitius*). Freshwater populations of both species evolved similar phenotypic traits after marine ancestors independently colonised freshwater environments. A highly resolved phylogeny is a prerequisite for untangling the processes that have shaped the underlying genomic divergence, including natural selection and population demographic history. Therefore, my thesis begins by resolving the worldwide phylogenetic relationships and demographic history of both focal species, using state of the art phylogenomic analyses. The results indicate that extant three-spined stickleback populations originated from the Eastern Pacific in the late Pleistocene, and the Atlantic populations were colonised from the Pacific ancestors via the Arctic Ocean. In contrast, nine-spined sticklebacks have a more ancient history, diversifying in the late Pliocene, and their current distribution is the result of multiple waves of trans-Arctic colonisation from the Far East, with several divergent lineages having evolved across their geographic range.

The thesis then moves on to investigate the genetic basis of parallel freshwater adaptation in each of the two species, using the information gained in the previous chapter to set up specific hypotheses and define simulation parameters. For three-spined sticklebacks, the level of parallel evolution at the genotype level was 10 times higher among the freshwater populations in the ancestral Eastern Pacific region than

anywhere else in the world. Empirical data and simulations demonstrate that these patterns are determined by a reduction in standing genetic variation outside the ancestral Eastern Pacific region, a result that can be explained by the demographic history of the species. A comparison of the two species revealed fundamental differences in the way standing genetic variation – the raw material upon which selection acts – is distributed among populations. This was exemplified by 2-fold higher degree of genetic structuring and 23-fold stronger isolation-by-distance in ninethan in three-spined sticklebacks. Conversely, the proportion of genetic parallelism in three-spined stickleback is 123.4 times greater than the nine-spined stickleback.

Taken together, the thesis resolved the phylogenetic affinities and demographic history of stickleback fishes using state-of-art methods and a global sampling strategy. Based on this knowledge, the thesis further uncovered profound heterogeneity in the repeatability of evolution within and between the two model species in response to freshwater colonisation. Hence, the two stickleback species with their contrasting demographic and evolutionary histories constitute a model system to study how differences in the distribution of standing genetic variation can influence the predictability of evolution.

Introduction

Phylogenomics

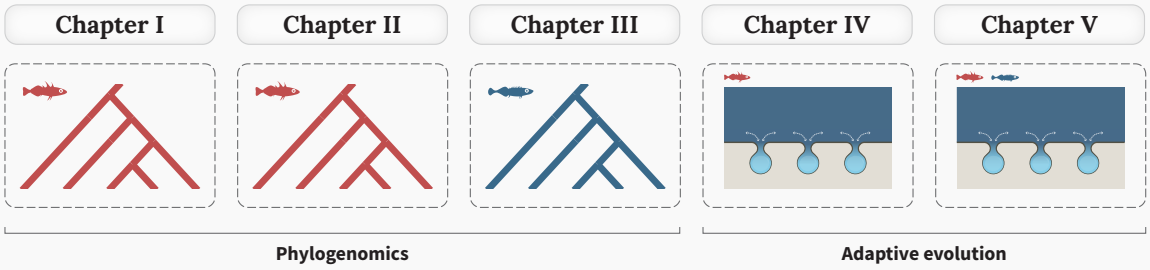
A well-resolved phylogeny among taxa of interest provides the evolutionary background on which we formulate and test rigorous ecological and evolutionary hypotheses (Delsuc et al. 2005). For instance, phylogenies are needed to gain insights into character homology and to identify the evolutionary mechanisms through which similar phenotypes are likely to evolve in distinct evolutionary lineages (Elmer & Meyer 2011; McCune & Schimenti 2012; Schluter 2000).

During the past two decades, genome-wide data has become increasingly available due to decreasing costs, increased computational power and the development of novel analytical tools, ultimately transforming molecular phylogenetics (Philippe et al. 2005; Telford et al. 2015). While during the 1990s it was very common to reconstruct phylogenies from short fragments of mtDNA, reconstruction of phylogenies using thousands of genes from hundreds of individuals is currently within the reach of any average-sized lab (Bleidorn 2017). This transformation has been enabled by the rapid development of high-throughput sequencing technologies (next-generation sequencing, NGS; Davey et al. 2011), which allows a large number of genetic markers to be obtained with deep genome-wide coverage in a cost-effective manner. The most common sequencing approaches used in ecology and evolutionary biology are whole genome sequencing (WGS; Davey et al. 2011) and restriction site-associated DNA sequencing (RAD-seq; Hohenlohe et al. 2010), which target the entire genome or a random subset (usually 1-10%), respectively.

Analysing data from a large number of independently segregating loci across the entire genome presents novel and significant challenges (Kapli et al. 2020). This is because various parts of the genome may reflect different evolutionary histories, resulting in an incongruence between the species tree and individual gene trees (Edwards et al. 2016; Nichols 2001). Such incongruence is a result of different evolutionary processes, mainly incomplete lineage sorting (ILS, **Box 2**), gene flow and gene transfer (Mallo & Posada 2016).

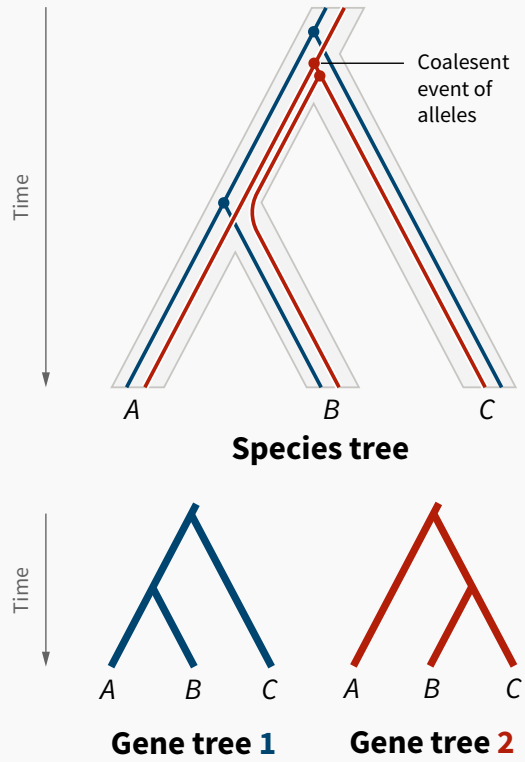
For recently diverged taxa, ILS is often assumed to be the main cause of discordance between individual gene trees and the species tree (**Box 2**). The multi-species coalescent (MSC) model can be applied to account for ILS

Box 1. Structure of the thesis



The main topics and species studied are presented. Red fish represent the three-spined stickleback (*Gasterosteus aculeatus*); blue fish represent the nine-spined stickleback (*Pungitius pungitius*).

Box 2. Incomplete lineage sorting



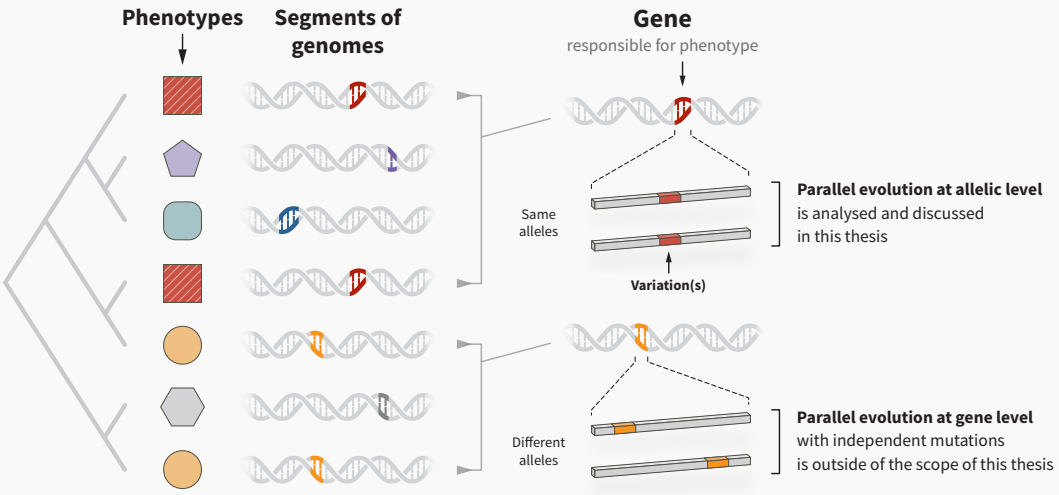
The figure illustrates a gene-tree-species-tree discordance caused by incomplete lineage sorting (ILS). ILS is the scenario where the most recent common ancestor for a given gene of the two species precedes speciation time, also known as deep coalescence (Tiley et al. 2020; Mallo & Posada 2016). This phenomenon is common, particularly when speciation or diversification events happened on a short time scale causing short branches connecting taxa. In this thesis, the inferred population trees are similar to the species trees mentioned above, but consider the history of conspecific populations. ILS can be modelled using the multi-species coalescent (MSC) model (Rannala & Yang 2003), which reconstructs gene trees separately and infers species tree based on the probability distribution of gene trees. Hence, MSC-based methods are believed to be superior to other methods in recovering true species trees among recently divergent taxa.

(DeGiorgio & Degnan 2009; Rannala & Yang 2003). In contrast to the common practice of concatenating all genes into one supermatrix from which the phylogenetic tree is inferred, the MSC model reconstructs species tree through integration over separate gene trees (Xu & Yang 2016). The implementation of the MSC model in a Bayesian framework is considered to be the best approach to solve the ILS issue (Leaché & Rannala 2011; Thawornwattana et al. 2018). This approach is based on Bayesian inference where the best phylogeny is searched for by estimating the statistical distributions of simulated parameters (topology, branch length, etc.) derived from a Markov chain Monte Carlo (MCMC) algorithm (Rannala & Yang 1996). However, the MSC model in a Bayesian framework is computationally highly demanding, and for this reason, application of the MSC method sometimes requires simplified models to reduce computational demand (e.g. Stange et al. 2018).

Parallel evolution

When multiple independent populations adapt to similar environments, they may evolve similar phenotypes in response to similar selective pressures. These parallel phenotypic changes are often referred to as parallel or convergent evolution, depending on whether the similar phenotypes arise via the same (parallel) or different (convergent) developmental pathways and genetic mechanisms (Futuyma 1998). Throughout this thesis, parallel evolution is strictly defined as parallel phenotypic changes arising from the same genetic mechanism (i.e. selection acting on the same alleles; See **Box 3**).

Box 3. Parallel evolution as defined in this thesis



The terms "parallel" and "convergent" evolution have various definitions in the literature. The traditional definitions in the context of phenotypic evolution is anchored to phylogeny/relatedness (Futuyma 1998): parallelism occurs when the phenotypic similarities have arisen in closely related taxa (e.g. loss of lateral plates in different stickleback populations), whereas phenotypic similarity among distantly related taxa (e.g. wings of birds and bats) is considered as convergence. However, an alternative definition is based on the molecular mechanisms through which phenotypic similarities evolve (Elmer & Meyer 2011; Rosenblum et al. 2014). In this definition, parallelism and convergence refer to the evolution of similar phenotypes via shared or distinct molecular mechanisms, respectively. Other definitions exist (see review in Rosenblum et al. 2014) and there is no consensus, as all definitions have their limitations (Haas & Simpson 1946). For instance, the criteria for differentiating "closely" from "distantly" related taxa is vague, and there are many hierarchical levels of genetic similarity at the molecular level (e.g. allele, gene, network, pathway, function; Conte et al. 2012; Rosenblum et al. 2014).

In this thesis, the term "parallel evolution" is used to refer both to phenotypic and genetic similarity in closely related, intraspecific taxa, always specifying whether I am referring to the phenotypic or genetic level. Moreover, genetic parallel evolution (or "genetic parallelism") is used to refer to parallel changes at the allelic level (i.e. selection repeatedly acting on the same variants). An example of parallel evolution at the nucleotide level is given in **Box 4**.

In the wild, parallel phenotypic changes are relatively common (e.g. Stern 2013; Rosenblum et al. 2014) and considered to be strong evidence for the action of natural selection, because the repeated evolution of similar phenotypes in response to similar ecological conditions is unlikely to occur by chance (Harvey & Pagel 1991). Many compelling cases of parallel phenotypic changes in various species, such as cichlid fishes and butterflies, can be found from recent reviews (e.g. Stern 2013; Bolnick et al. 2018). An iconic example of parallel evolution is the repeated reduction of skeletal armour in response to freshwater colonisation from marine ancestors of three-spined sticklebacks; the same low-plated phenotypes evolved in independent populations via repeated selection

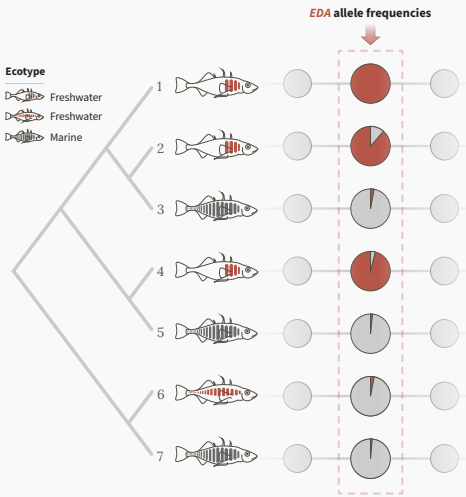
on the same allele (Bell & Foster 1994; Colosimo et al. 2005) (**Box 4**).

Understanding parallel evolution at both phenotypic and molecular levels can improve our ability to predict adaptive evolution (Orr 2005; Sackton & Clark 2019). In recent decades, research has focused on exploring the genomic mechanisms underlying parallel phenotypic changes in order to understand the extent to which the phenotypic similarities across taxa are driven by the same genetic mechanisms (Conte et al. 2012). A common approach to investigate genetic parallelism is to perform F_{ST} based genome scans among replicate pairs of populations, seeking overlapping outliers that show signatures of selection in multiple population pairs (reviewed by Fraser & Whiting 2019; **Box 5**). Unsupervised genome scan approaches (e.g., Luu et al. 2017; Jones et al. 2012; Kemppainen et al. 2015) are also increasingly adopted in the analyses of genetic parallelism because of their capability to uncover evolutionary phenomena without a priori population classification (**Box 5**).

Parallel evolution can be achieved via selection on shared genetic variation segregating in the common ancestor (i.e. standing genetic variation, SGV) or via repeated de novo mutations in segregating lineages (Barrett & Schluter 2008). Adaptation from SGV can lead to rapid evolution in response to new environments, and is seen as the principal source of variation fuelling parallel evolution across populations (Barrett & Schluter 2008; Schluter & Conte 2009; Thompson et al. 2019). Empirical cases of rapid parallel evolution from SGV have been observed to occur over just a few decades (Lescak et al. 2015; Marques et al. 2018).

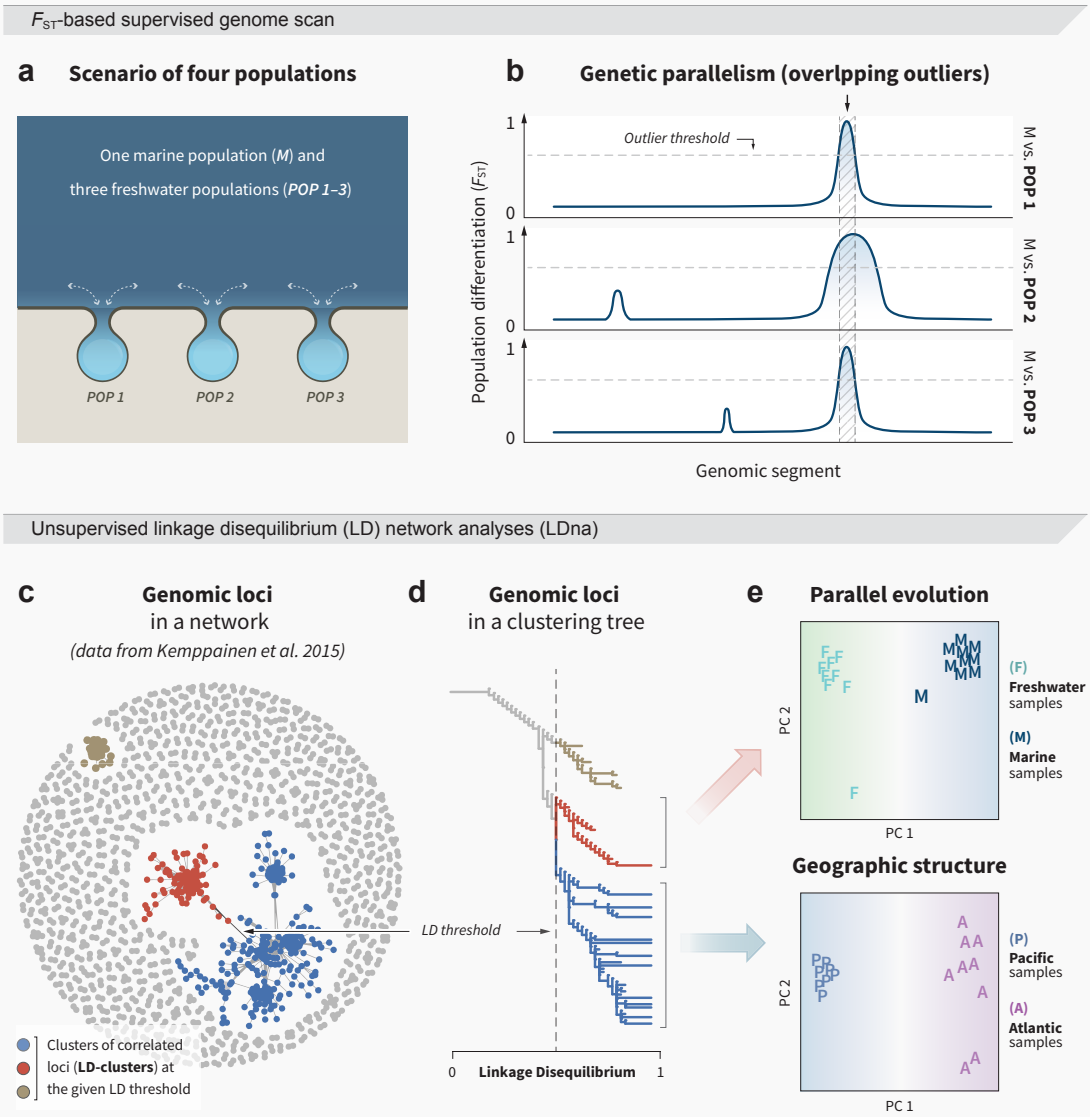
The probability of parallel evolution is determined by the similarity of selective optima, the availability of a common pool of adaptive alleles (SGV) and the genetic architecture of the trait in question (Rosenblum et al. 2014). In turn, SGV is a product of the demographic history of a set of populations, as well as their history of selection. While the effect of the variance in selective optima and the genetic constraints on the probability of parallel evolution have received much attention (e.g. Bailey et al. 2015; Colosimo et al. 2005; Jones et al. 2012; Morales et al. 2019; Rennison et al. 2020; Stuart et al.

Box 4. Parallel evolution in lateral plate phenotypes in three-spined sticklebacks.



THE EVOLUTION of ecologically divergent lateral plate phenotypes in three-spined sticklebacks (*G. aculeatus*) serves as an example of genetic parallelism at the allelic level in the wild. The schematic illustration shows the reduction in the number of lateral plates in sticklebacks in response to freshwater adaptation from the marine habitat. This repeated phenotypic evolution of low-plated phenotypes across freshwater populations occurred through repeated selection of Ectodysplasin (*EDA*) alleles derived from an ancestral haplotype in the marine populations (Colosimo et al. 2005). This is illustrated in the accompanying figure where parallel evolution occurs in the fish populations (1), (2) and (4) due to the reused low-plated *EDA* alleles. When the low-plated allele frequency (red colour in pies) is low or missing, the fish might evolve a functionally equivalent freshwater phenotype (e.g. 'small-plated' fish population [6] in the figure) through alternative genetic pathway(s) instead of the *EDA* gene (Leinonen et al. 2012).

Box 5. Detecting genetic parallelism using supervised and unsupervised genome scans in the thesis.



Schematic representation of genome scan approaches to detect parallel genetic changes: F_{ST}-based supervised genome scan (a-b) and unsupervised linkage disequilibrium (LD) network analyses (LDna, c-e). (a) A demographic scenario of three independent freshwater populations originating from the same ancestral marine population. (b) Genetic parallelism (overlapping outliers) across three freshwater populations identified using pairs of F_{ST} genome scans. (c) A schematic clustering tree of pairwise LD values among genomic loci in the population genomic data of Kemppainen et al. (2015). (d) Network visualisation showing all links between loci in a population genomic dataset at a given LD threshold. (e) Evolutionary phenomenon of LD-clusters revealed by PCA, reflecting parallel genetic evolution (top) and geographic structure (bottom).

GENOME SCANS among replicate population pairs are a useful approach to investigate parallel evolution, as they are easily implemented even in the context of non-model species with average sample sizes (i.e. N > 10) and without genetic crosses (Whiting & Fraser 2020). All genome scans aim to identify loci that display stronger differentiation than expected under a neutral null model of divergence (Butlin 2010). Among populations, allelic differentiation (F_{ST}; Weir & Cockerham 1984) is the most commonly used statistic in outlier analyses (reviewed by Fraser & Whiting 2019). This approach is "supervised", in the sense that it requires a priori classification of populations with divergent adaptations (i.e., ecotypes; e.g., marine *versus* independent freshwater stickleback populations; [a] in the figure). The presence of common genomic outliers among replicate pairs of populations under divergent selection are thus taken as evidence of genetic parallelism (b).

The supervised approach might lead to biases in outlier detection when 1) populations in a continuous sampling scheme are defined improperly (Waples & Gaggiotti 2006), 2) populations contain admixed individuals (Lotterhos & Whitlock 2015; Luu et al. 2017), or 3) there are not enough samples at the population level to obtain precise estimates of allele frequencies. On the contrary, unsupervised genome scan approaches allow the analysis of a population genomic dataset without a priori population classification and lower levels of replication at the population level. Unsupervised genome scan approaches include the program pcadapt (Luu et al. 2017), which searches for outliers based on PCA and population structure, as well as the method of self-organizing map-based iterative Hidden Markov Model (SOM/HMM) used in Jones et al. (2012) to identify genetic parallelism across individuals with shared phylogenetic signals.

To identify genetic parallelism from population genomic datasets, this thesis applies another unsupervised approach – Linkage disequilibrium (LD) network analyses (LDna) – developed by Kemppainen et al. (2015). LDna adopts network analytical methods on the pattern of pairwise LD (Hill & Robertson 1968; Barton 2011) between loci across the genome (e.g., Fig. **Box 5c**), in which LD refers to the non-random association of alleles between pairs of loci. In LDna, a hierarchical (single linkage) clustering algorithm is used to cluster loci sharing high LD values (r^2; $0<r^2<1$, where 0 means that alleles are totally independent from each other) that aims to identify sets of loci connected by high LD (Fig. **Box 5d**). Since LD is a sensitive indicator of many evolutionary phenomena, the resulting LD-clusters (sets of clustered loci) can be attributable to population demographic history (Fig. **Box 5f**), chromosomal inversions (see example in Fig. 5e), local adaptation and parallel adaptation (Fig. **Box 5e**). When a PCA based on loci within a LD-cluster is performed, individuals are expected to group based on their demographic history (Fig. **Box 5f**), karyotype (in the case of chromosomal inversion) or ecotype (in the case of parallel evolution).

2017; Thompson et al. 2019), the impact of the variation in access to SGV remains much less studied (but see: Leinonen et al. 2012; Kemppainen et al. 2020; Ralph & Coop 2010; Ralph & Coop 2015a, b). The demographic history of a species (e.g., effective population size [Ne], gene flow and time of divergence among populations) determines the distribution of SGV within and across its populations, and thus may affect the probability of parallel genetic evolution (Rosenblum et al. 2014; Kemppainen et al. 2020).

The study systems

The focal species of this thesis are two co-distributed stickleback fishes in the family Gasterosteidae: the three-spined stickleback (*Gasterosteus aculeatus*) and the nine-spined stickleback (*Pungitius pungitius*; Fig. 1). The three-spined stickleback has been one of the most important model species for ecological and evolutionary biology research for decades (Bell & Foster 1994; Gibson 2005; Hendry et al. 2013; Östlund-Nilsson et al. 2006). Although serving as a model for some behavioural research in the past (Herczeg et al. 2009b; Herczeg et al. 2009c; Morris 1951, 1955), the nine-spined stickleback has been utilised in evolutionary biology research much less frequently (Merilä 2013). The two species diverged about 25 million years ago (Mya; Varadharajan et al. 2019), and share a similar, yet not identical, circumpolar distribution range across the northern hemisphere (Wootton 1976). Both species share similarities in size (typically <100 mm), ecological and life history traits, breeding habits, as well as in their morphological and behavioural characteristics (Baker 1994; DeFaveri et al. 2014; DeFaveri et al. 2013; McLennan & Mattern 2001; Wootton 1976; Wootton 1984). Despite these similarities, the three-spined stickleback is typically more abundant

in pelagic and coastal marine habitats than the nine-spined stickleback (Wootton 1976, 1984; DeFaveri et al. 2012; Ojaveer et al. 2003).

The *Pungitius* sticklebacks are more diversified than the *Gasterosteus* sticklebacks. The genus *Gasterosteus* consists of three taxonomically valid species (viz. *G. aculeatus*, *G. wheatlandi*, *G. japonicus*; Eschmeyer et al. 2017), while there are at least seven taxonomically valid species in the genus *Pungitius* (Eschmeyer et al. 2017; Takahashi et al. 2016). The level of genetic differentiation among populations in the genus *Pungitius* is much stronger than that among *Gasterosteus* populations (DeFaveri et al. 2012; Merilä 2013, 2014; but see Raeymaekers et al. 2017). Therefore, the pool of SGV across the distribution range of *Pungitius* species is likely to be more heterogeneous than that of the *Gasterosteus* species.

Three-spined stickleback (*G. aculeatus*)

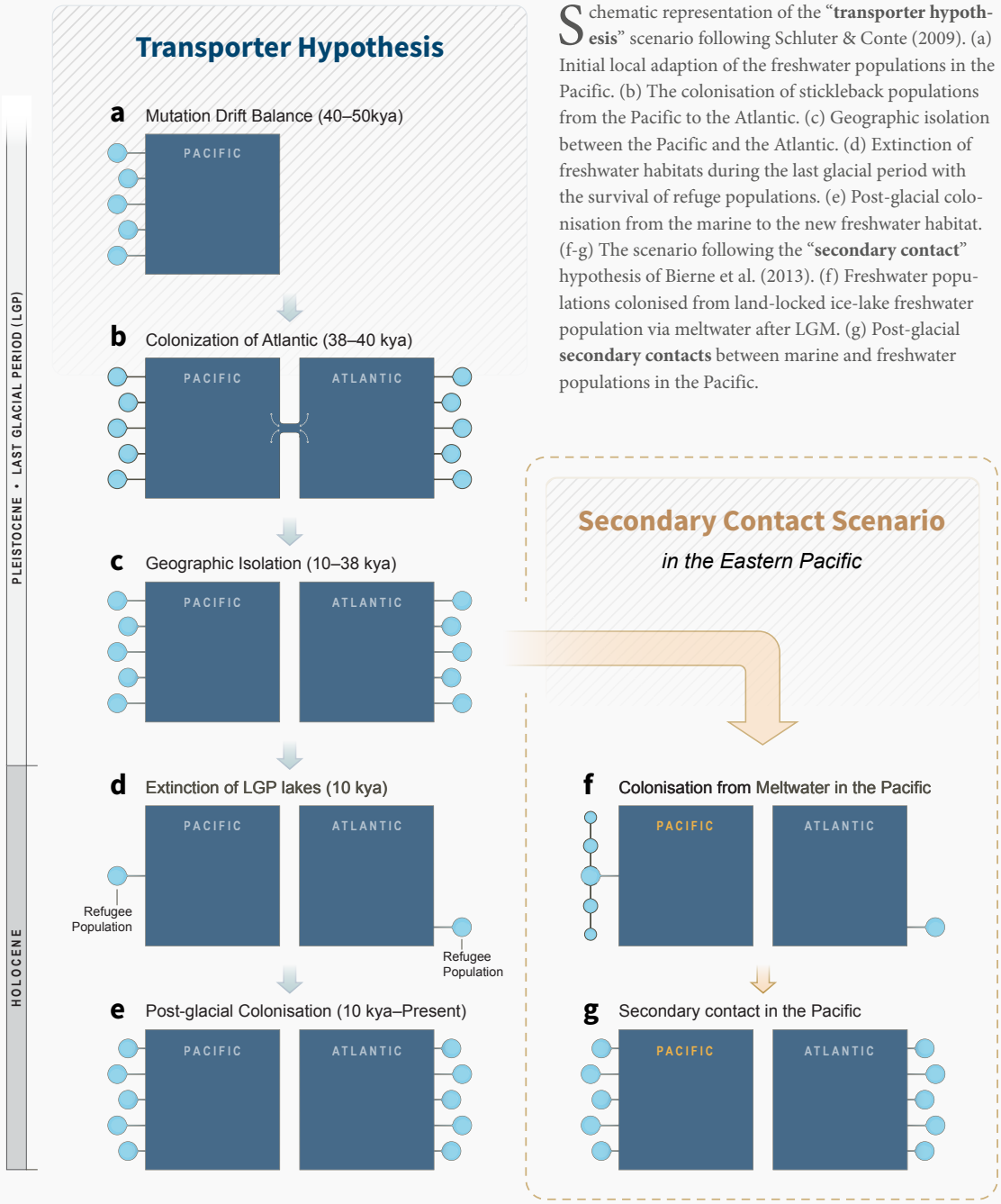
Nine-spined stickleback (*P. Pungitius*)

Figure 1 | The study species. The two main study species used as model in this thesis work (photos by Petri Kuokka and Bohao Fang).

Repeated evolution of similar phenotypic traits in similar habitats has been observed in both species, in terms of morphology, behaviour and life history (Leinonen et al. 2006; Östlund-Nilsson et al. 2006; Herczeg et al. 2010; Hohenlohe & Magalhaes 2019; Schluter et al. 2010; Merilä 2013; Kemppainen et al. 2020). For instance, three-spined sticklebacks have evolved a reduced pelvic apparatus and lateral plate numbers in numerous independently colonised freshwater environments (Chan et al. 2010; Colosimo et al. 2005; Cresko et al. 2004; Shapiro et al. 2004). Likewise, nine-spined sticklebacks have repeatedly evolved reduced body armour (Herczeg et al. 2010), including pelvic reduction (Kemppainen et al. 2020; Shapiro et al. 2006; Shikano et al. 2013), as well as larger body size in isolated ponds compared to their marine conspecifics (Herczeg et al. 2009a; Karhunen et al. 2014).

With access to genomic resources and new analytical tools, the genetic basis of sticklebacks' parallel evolution has been studied extensively. The most well-known case is the lateral plate reduction in the three-spined sticklebacks, which was found to be controlled by a major gene of large effect (Ectodysplasin, *EDA*; Colosimo et al. 2005; **Box 4**). Furthermore, a study using whole genome sequences of 20 three-spined sticklebacks sampled across the world identified ~200 genomic regions showing signatures of parallel freshwater adaptation (Jones et al. 2012). In contrast, it has been suggested that the repeated evolution of similar phenotypes in nine-spined sticklebacks is more likely to be grounded on more heterogeneous genetic mechanisms than that of three-spined sticklebacks (Merilä 2013). The reasoning behind this expectation is that populations of nine-spined

Box 6. Possible demographic scenarios of freshwater adaptation in three-spined sticklebacks.



Schematic representation of the "**transporter hypothesis**" scenario following Schluter & Conte (2009). (a) Initial local adaption of the freshwater populations in the Pacific. (b) The colonisation of stickleback populations from the Pacific to the Atlantic. (c) Geographic isolation between the Pacific and the Atlantic. (d) Extinction of freshwater habitats during the last glacial period with the survival of refuge populations. (e) Post-glacial colonisation from the marine to the new freshwater habitat. (f-g) The scenario following the "**secondary contact**" hypothesis of Bierne et al. (2013). (f) Freshwater populations colonised from land-locked ice-lake freshwater population via meltwater after LGM. (g) Post-glacial **secondary contacts** between marine and freshwater populations in the Pacific.

Schluter and Conte (2009) proposed the **"transporter hypothesis"** to explain rapid parallel evolution observed in three-spined stickleback populations from marine to freshwater. The hypothesis postulated that freshwater adaptation was facilitated by repeated selection on freshwater adapted alleles that are maintained as SGV in ancestral marine populations in low frequency.

An alternative explanation for the observed parallel ecotypic divergence in three-spined sticklebacks is the scenario of **"secondary contact"** proposed by Bierne et al. (2013). Using simulations, Bierne et al. (2013) suggested that the same ecological divergent genomic pattern could be achieved when the population range expansion initiated in allopatry, followed by gene flow upon secondary contact

between marine and freshwater habitats. From the genetic perspective, the gene flow between ecotypes after range expansion would erase the past differentiation in neutral loci but not for genomic regions under divergent selection, and thus shape the exact same pattern attributable to the ecological speciation explained by the **"transporter hypothesis"** (Bierne et al. 2013).

Following the above two hypotheses, the figure depicts the possible demographic scenarios of freshwater adaptation in three-spined sticklebacks to explain the high levels of parallel genetic evolution in the Eastern Pacific region (see **results and discussion**).

stickleback are older and genetically more structured (subject to more genetic drift) than those of three-spined sticklebacks (DeFaveri et al. 2012). This pattern is in turn expected to reduce the pool of shared SGV among populations, lowering the probability of parallel evolution. There is some evidence to suggest that this is indeed the case (Kemppainen et al. 2020), but comparative genomic studies of the two co-distributed stickleback species utilising broad geographic and genomic sampling are still lacking.

It is widely acknowledged that three-spined sticklebacks exhibit high levels of parallel evolution as exemplified by numerous genomic regions consistently differentiating marine and freshwater ecotypes (Hohenlohe & Magalhaes 2019). However, the historical focus in previous studies has been on the Eastern Pacific region (Hohenlohe et al. 2010; Jones et al. 2012; Chan et al. 2010; Colosimo et al. 2005; Hohenlohe & Magalhaes 2019; Nelson & Cresko 2018). Although sampling in two studies has covered a larger geographic range across the Pacific and Atlantic Oceans (Jones et al. 2012; DeFaveri et al. 2011), the Eastern Pacific samples still constituted over half of the samples in the study by Jones et al. (2012). In fact, recent studies focusing on populations from the Atlantic region have indicated much more heterogeneous ecotype differentiation with different and relatively limited genetic parallelism across studied populations (Ferchaud & Hansen 2016; Liu et al. 2018; Pujolar et al. 2017; Terekhanova et al. 2019; Terekhanova et al. 2014). For instance, using RAD-seq, Ferchaud & Hansen (2016) did not find any consistently differentiated genomic regions between all marine–freshwater population pairs in Denmark. Therefore, in order to quantify the extent and geographic heterogeneity of parallel evolution in the three-spined stickleback supermodel, a genome-wide analyses based on the comprehensive global sampling is needed.

Aims of this thesis

The broader aim of this thesis was to gain insights into the evolutionary mechanisms influencing the repeatability (i.e. predictability) of local adaptation in populations adapting to similar environments. More specifically, I wanted to investigate if and how the phylogenetic and demographic history of populations constrains the probability of parallel evolution. To this end, I studied the phylogenetic and demographic history and incidence of parallel adaptive evolution in both target species. (**Box 1**; **Table 1**). First, I constructed the worldwide phylogeny of the three-spined stickleback (**Chapter I and II**), and that of the nine-spined stickleback (**Chapter III**). Second, I investigated the patterns and underlying causes of heterogeneous parallel evolution in three-spined stickleback marine and freshwater populations (**Chapter IV**), as well as that between three- and nine-spined sticklebacks (**Chapter V**).

In **Chapter I**, the aim was to reconstruct the phylogenetic relationships and colonisation routes among worldwide three-spined stickleback populations. Since all previous studies had used either limited geographic sampling and/or limited number of marker genes, the evolutionary relationships and colonisation history of its populations have remained poorly resolved. The study sought to infer a robust and exhaustive phylogeny by using genome-wide SNPs and comprehensive sampling of populations spanning the species' entire distribution range.

In **Chapter II** the aim was to estimate divergence times among major three-spined stickleback lineages based on the phylogenetic topologies recovered in **Chapter I**, and to explore if and how ILS affects the divergence time estimates. The motivation for this study was provided by the findings of Stange et al. (2018), which indicated that the site concatenation approach used also in **Chapter I** might bias divergence time inference in the presence of ILS, and that such bias could be accounted for by the MSC model. By applying both a concatenation approach and the MSC method with multiple calibration schemes, the study sought to establish a robust timeline for the diversification of worldwide stickleback lineages and to provide a case study to illustrate how different analytical frameworks and calibration strategies affect divergence time estimates.

In **Chapter III**, I switched focus to the genus *Pungitius* and its most widespread member, the nine-spined stickleback (*P. pungitius*). The main aims were to clarify the phylogenomic relationships of *Pungitius* species and populations, and to investigate the proportion of the genome subjected to introgression among *Pungitius* taxa. Moreover, based on the phylogenomic hypothesis and sequence divergence analyses, I sought to decipher the taxonomic validity of *Pungitius* taxa which have been subject to long-standing controversy due to conflicting evidence from morphological and mitochondrial DNA analyses.

In **Chapter IV**, I studied parallel evolution in the three-spined stickleback. Using a large genomic data set with comprehensive global sampling, I explored genetic parallelism in marine–freshwater differentiation at different geographic scales, considering the phylogeographic affinities and colonisation history of *G. aculeatus* populations revealed in **Chapters I and II**. The aim was to examine whether the degree of parallel evolution is as pervasive as suggested in earlier studies, or whether there is geographical heterogeneity in the degree of parallelism that could be attributable to the complex demographic history of this species.

Table 1. Summary of the main objectives, methods, results and implications of individual chapters included in the thesis.

Chap.	Main objectives	Main methods	Main results	Implications
I	Reconstruction of the phylogenetic relationships and colonisation histories among worldwide *G. aculeatus* populations.	Bayesian coalescent analyses based on concatenation of thousands of genome-wide single-nucleotide polymorphisms (SNPs).	Robust phylogeny and reconstruction of colonisation history of worldwide *G. aculeatus* populations.	Extant three-spined sticklebacks share a very recent ancestry and have colonised Atlantic Ocean from the Eastern Pacific in the Late Pleistocene, far more recently than previously thought.
II	Estimation of divergence times among major clades of three-spined sticklebacks, and testing whether incomplete lineage sorting (ILS) may have biased estimates of divergence times in **Chapter I**.	Phylogenetic methods using a multi-species (MSC) model under a Bayesian framework, and the concatenation method of **Chapter I**.	Robust divergence time estimates of three-spined sticklebacks, incorporating the uncertainties in different calibration schemes and phylogenetic methods.	Both calibration schemes and the multi-species coalescent model impact the divergence time estimates. Multiple analytical frameworks are advocated in divergence time estimation.
III	Uncover the evolutionary relationships among different *Pungitius* species and populations globally, as well as study the prevalence and extent of introgression among recognized species.	MSC and maximum likelihood-based phylogenetic inferences and introgression detection with D-statistic and D_{FOIL}-statistics tests.	Evolutionary relationships within the *Pungitius* complex resolved. Taxonomic validity of different taxa clarified. Evidence for frequent hybridization among taxa.	The utility of mitochondrial markers in the study of evolutionary relationships among taxa is limited, particularly when hybridization and introgression have occurred.
IV	Quantify the pervasiveness (or lack thereof) of genetic parallelism underlying freshwater adaptation in three-spined sticklebacks on a global scale.	Detection of marine-freshwater differentiated genomic regions using Linkage disequilibrium (LD) network analyses (LDna) and F_{ST}-based genome scans.	Three-spined sticklebacks exhibit strikingly higher levels of genetic parallelism in the ancestral Eastern Pacific region than anywhere else in the world in response to freshwater colonisation.	Significantly reduced ancestral standing genetic variation (SGV) outside the Eastern Pacific region suggests that demographic history has an important role in shaping evolutionary adaption and the likelihood of parallel evolution.
V	Compare the distribution of genetic variation within and among populations of the two stickleback species. Compare levels and patterns of the genetic parallelism in response freshwater colonisation among the two species.	Comparative population genetic analyses, comparative phylogenomic analyses and comparative LDna analyses.	The nine-spined stickleback harbours more heterogeneous pools of SGV and much lower levels of genetic parallelism than the three-spined stickleback.	The distribution of SGV – attributable to differences in species' demographic and evolutionary histories – influences the predictability of evolution.

Particularly, I wanted to test the hypothesis that the loss of standing genetic variation that has occurred since the species colonised areas outside of the ancestral Eastern Pacific region has reduced the probability of parallel evolution.

In **Chapter V**, I compared genetic diversity, evolutionary history and the degree of marine–freshwater genetic parallelism between the three- and nine-spined sticklebacks. In particular, I sought to identify if differences in the distribution of SGV can explain the differences in incidence of parallel evolution between the two species. Here, I hoped to shed light on the factors that shape the predictability of evolutionary adaptation to similar selection pressures by comparing the degree of genetic parallelism in marine–freshwater differentiation in the two species and relating it to factors differentiating the two species.

Materials and Methods

In the following section, I will briefly introduce the materials and methods used in the five chapters of the thesis. Detailed descriptions of the methods and bioinformatics pipelines used are available from the methods sections of the individual chapters.

The study species and sampling

This thesis focused mainly on two species: the three-spined stickleback (*G. aculeatus*; **Chapter I, II, IV and V**) and the nine-spined stickleback (*P. pungitius*; **Chapter III and V**). The fish samples used in this thesis were collected with seine nets, minnow traps or electrofishing by research group members or collaborators.

The three-spined stickleback samples used in **Chapter I** were collected throughout the species' geographic range across the Pacific and Atlantic Oceans, including samples from the Mediterranean basin (Fig. 2a). Both marine and freshwater ecotypes were included. Most samples used in **Chapter I** were also used in the following chapters. **Chapter II** used a subsample of populations from **Chapter I**. These populations were chosen to represent major lineages covering the entire species' geographic range. Two *G. nipponicus* individuals were used in **Chapters I and II** as outgroups for phylogenetic inferences. **Chapter IV** used samples from **Chapter I**, excluding the populations from the Mediterranean basin (three-spined sticklebacks in the Mediterranean area are restricted to freshwater habitats). In addition, **Chapter IV** also incorporated some newly sequenced samples as well as samples from public databases. **Chapter V** used the same three-spined stickleback samples used in **Chapter IV**.

Chapter III used samples of nine-spined stickleback (Fig. 2b) and six other *Pungitius* species, including *P. pungitius*, *P. laevis, P. platygaster, P. hellenicus, P. sinensis, P. tymensis* and *P. kaibarae*. Their sampling map can be found in **Chapter III**. In **Chapter V**, I used the samples from **Chapter V** as well as some new samples sequenced specifically for this study. In this chapter, I aimed to match the geographic sampling of nine-spined and three-spined sticklebacks as closely as possible.

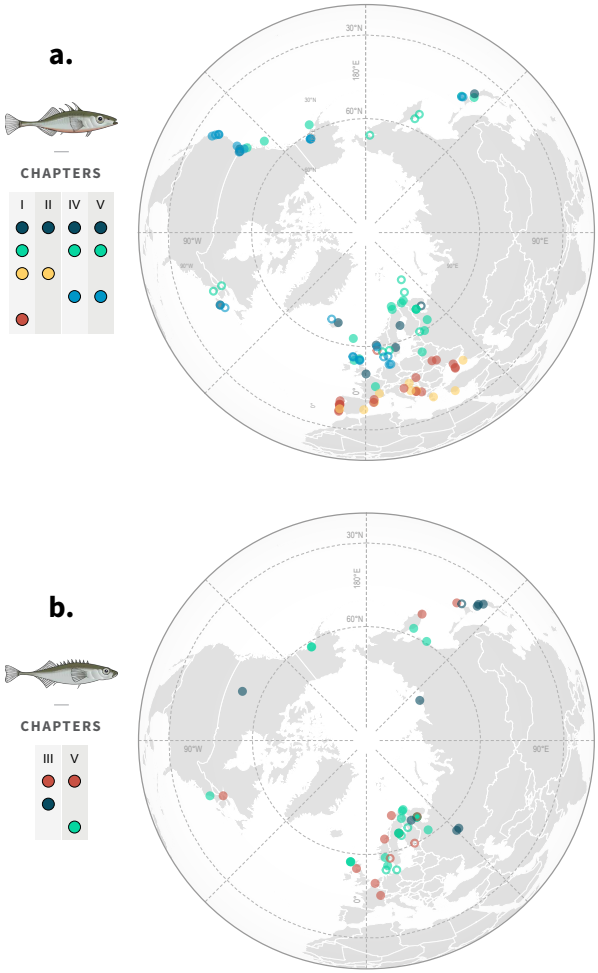


Figure 2 | Sampling map of the study species. (a) Three-spined stickleback (*G. aculeatus*) populations used in **Chapters I, II, IV and V**. (b) Nine-spined stickleback (*P. pungitius*) populations used in **Chapters III and V**. Other *Pungitius* species in the genus are not displayed, but their sampling localities can be found in **Chapter III**. Solid circle, freshwater populations; hollow circle, marine populations.

Sequencing, genotyping and genotype likelihood estimation

In the thesis, all chapters used genome-wide SNP data sequenced by RAD-seq or WGS approaches. Protocols of DNA extraction, library preparation and sequencing are given in the respective chapters. In **Chapters I, II and III**, the phylogenomic and population genetic analyses were conducted based on genotypes. In **Chapter IV and V**, which investigated marine–freshwater genetic parallelism, the population genomic data sets were prepared in the form of genotype likelihoods to account for heterogeneous quality of the sequence data from several sources.

The genotyping was conducted using standard bioinformatic pipelines. The raw sequencing data of the two species were first checked with FastQC (Gordon & Hannon 2010) and mapped to their respective reference genomes using BWA (Li and Durbin 2010). The mapped reads were used for calling variants (SNPs) with SAMtools and BCFtools (**Chapter III**; Li 2011) or ipyrad (**Chapter I and III**; Eaton 2014). Genotype filtering was conducted in VCFtools (**Chapter I, II, III, IV and V**; Danecek et al. 2011). In **Chapter IV and V**, genotype likelihoods estimation and their quality control were obtained from mapped reads with the program suite ANGSD (Korneliussen et al. 2014). Sex chromosomes of both species were excluded in the analyses.

Phylogenomic analyses

The phylogenetic analyses from **Chapters I, II and III** were performed with three different methods depending on the data sets used and the purpose of the analyses. The first method was the maximum-likelihood (ML) analysis implemented in RAXML (Stamatakis 2014) based on site supermatrices, which was able to handle large phylogenomic data. The second

method was the Bayesian coalescent analysis with a relaxed molecular clock model implemented in BEAST (Bouckaert et al. 2014) based on the concatenation of loci within a supermatrix. This approach has been shown to be less affected by ILS than the ML method (Lambert et al. 2015). The third method was the Bayesian phylogenetic inference based on the MSC model as implemented in SNAPP (Bouckaert et al. 2014). While this method can model ILS, it is also computationally demanding and currently limited to analyses of a few tens/hundreds of individuals with a few thousand SNPs (Zimmermann et al. 2014).

The coalescent-based phylogenies (concatenation and MSC methods) were applied with calibration points to infer divergence times; details regarding calibration strategies can be found in the respective chapters. The confidence of the inferred phylogenies was tested using bootstrapping for ML trees, and evaluated by the posterior probabilities for coalescent trees.

In **Chapters I and II**, I used a concatenation approach to explore the time-calibrated worldwide phylogeny of three-spined sticklebacks. The MSC model was used in **Chapters II, III and V** to infer the time-calibrated phylogenies of three- and nine-spined sticklebacks (or *Pungitius* taxa) while accounting for ILS. The ML method was adopted in **Chapter III** to provide a reference phylogeny for *Pungitius* species using a large SNP supermatrix.

Population genetic analyses

Genetic diversities (individual heterozygosity *H;* nucleotide diversity π, Nei & Li 1979; Watterson's theta θ, Watterson 1975) and genetic differentiation (F_{ST}, Weir & Cockerham 1984) were compared within (genetic diversity) and between (F_{ST}) populations of both species (**Chapter V**). To calculate *H,* I obtained the site frequency spectrum (SFS) for each individual and divided the number of segregating sites by the sum of the SFS. To calculate π and θ, per-site genetic diversities were firstly estimated within populations based on the SFS with ANGSD starting from genotype likelihoods, and then averaged across sites to obtain global values. Pairwise F_{ST} between populations and the global F_{ST} of each ecotype were estimated based on the genotypes called from ANGSD with the R packages hierfstat (Goudet 2005) and StAMPP (Pembleton et al. 2013). Genetic diversities and allelic differentiation were quantitatively compared between species with statistical tests by fitting generalized linear mixed-effects models (GLMMs) or by bootstrapping (**Chapter V**).

To evaluate and compare gene flow in the two focal species, Isolation-by-Distance (IBD) analyses were performed for each ecotype of the two species (**Chapter V**). The analyses were conducted by regressing pairwise genetic distances (linearized $F_{ST} = F_{ST}/(1 - F_{ST})$; Rousset 1997) against pairwise geographic distances between populations. To test and compare the significance and levels of IBD across species, the regressions were fitted with maximum-likelihood population effects (MLPE) models, accounting for non-independence of pairwise distances (Clarke et al. 2002).

Genomic introgression among *Pungitius* taxa was evaluated with D-statistics (known as the 'ABBA-BABA' test; Durand et al. 2011) and its extension known as D_{FOIL} statistics (a five-taxon test; Pease & Hahn 2015; **Chapter III**). The D-statistics and D_{FOIL} detect the admixed fraction of the genome by summarising the proportion of genomic data (SNPs) that is biased in respect to the topologies expected under a strict bifurcating evolutionary history. The latter test could also identify the introgression do-

nor and recipient lineages (i.e. direction of introgression). Their specific applications were stated in **Chapter III**.

Detection of genetic parallelism

To study the genetics of parallel evolution in response to freshwater colonisation in the two species, I assessed the patterns of genomic differentiation between marine and freshwater ecotypes in both species (**Chapter IV and V**). Two approaches were used for this purpose (**Box 5**). The first approach relies on supervised genome scans based on marine–freshwater allelic differentiation (F_{ST}). The second approach was the unsupervised Linkage Disequilibrium (LD) Network Analysis (LDna; **Box 5**).

For the F_{ST} analyses, the SNP-based F_{ST} between marine and freshwater ecotypes was calculated across the genome in respective species with ANGSD. In **Chapter IV**, the marine–freshwater F_{ST} of three-spined sticklebacks was analysed separately for different geographical regions (Eastern Pacific, Western Pacific and Atlantic). In **Chapter V**, the marine–freshwater F_{ST} was computed for Atlantic populations of three- and nine-spined sticklebacks, respectively (we had no marine nine-spined stickleback samples from the Pacific regions).

The second approach, LDna, is able to separate population genomic data into sets of highly correlated loci (LD-clusters) that reflect distinct evolutionary processes (**Box 5**). The association between the phylogenetic signal (inferred by Principal Component Analyses [PCA]) and marine–freshwater parallelism for each LD-cluster was tested by permutation. The resulting clusters of loci (LD-clusters) in LDna can be visualised in the form of networks where loci (nodes) are connected with LD values (edges) above given thresholds (**Box 5**). LDna was performed with custom R scripts provided in the individual chapters. In **Chapter IV**, LDna was applied to the data set of global three-spined stickleback populations for identifying and quantifying marine–freshwater genetic parallelism at the regional and global geographical scales. In **Chapter V**, LDna was applied to three- and nine-spined stickleback populations in the Atlantic region to explore and compare the levels of genetic parallelism between the two species.

Results and Discussion

Phylogenomics

Nine-spined sticklebacks were found to have a more ancient and complicated evolutionary history than three-spined sticklebacks. The time to the most common ancestor (TMRCA) of three-spined stickleback lineages could be traced back to the late Pleistocene, while that of nine-spined sticklebacks was much earlier in the late Pliocene. Extant Atlantic three-spined stickleback populations originated from a very recent colonisation from the Pacific just before the Last Glacial Maximum (LGM, ca. 20–26 thousand years ago, or Kya), and exhibit significant ILS among newly founded postglacial Northern European populations (**Chapter I**, **II and V**). In the case of the nine-spined stickleback, multiple trans-Arctic colonisations between the Pacific and Atlantic Oceans were identified, resulting in several divergent lineages across the species distribution range. For instance, the phylogenetically young European populations were split into two distinct lineages, interpreted to be a result of two independent trans-Arctic colonisations from the Far East and/or North America, all of which happened before the LGM (**Chapter III and V**). I briefly elaborate on the results and discussions of the phylogenomic analyses below.

Phylogenomics of three-spined sticklebacks

The phylogenomic analyses revealed three major clades among worldwide populations (**Chapter I**). These can be classified as a Trans-Pacific clade, a Southern European clade and a Trans-Atlantic clade (Fig. 3). The results of biogeographic analyses (**Chapter I**) and divergence time estimates (**Chapter I and II**) suggest the following colonisation histories: the extant three-spined stickleback populations originated from the Pacific Ocean in the Late Pleistocene (ca. 36.9–346.5 Kya, Fig. 3), and colonised the Atlantic through the Bering Sea and Arctic Ocean ca. 29.5–226.6 Kya (Fig. 3). This Atlantic lineage likely survived in Southern European refugia during glacial periods and recolonised Northern Europe and the Western Atlantic (North American east coast) following the end of the last glaciation (Fig. 3).

The following findings regarding phylogenetic relationships and colonisation history of three-spined sticklebacks are worth emphasising. First, the most ancestral populations resided in the Eastern Pacific region, from where three-spined sticklebacks colonised the rest of the world. This finding provided demographic evidence for my subsequent studies (**Chapters IV and V**), indicating that the Eastern Pacific is likely to harbour more SGV. Second, all contemporary three-spined stickleback populations share a very recent ancestry. The TMRCAs for all lineages inferred using coalescence-based methods in **Chapters I and II** were much more recent than previous estimates derived from mtDNA-based studies (Mäkinen & Merilä 2008; Orti et al. 1994) and the fossil evidence from the Atlantic basin (Bell & Foster 1994; Foster 1995). These findings confirmed that the Late Pliocene/Early Pleistocene Atlantic populations had gone extinct (Orti et al. 1994; Mäkinen & Merilä 2008) before the recent colonisation of the extant Atlantic populations. Third, the worldwide phylogeny resolved the decade-long controversy over the origin of the Black Sea populations. Some previous studies suggested that Black Sea populations originated from a recent invasion from the Mediterranean Sea (Mäkinen et al. 2006; Mäkinen & Merilä 2008) rather than by pre-Pleistocene colonisation from North-eastern Europe (Mün-

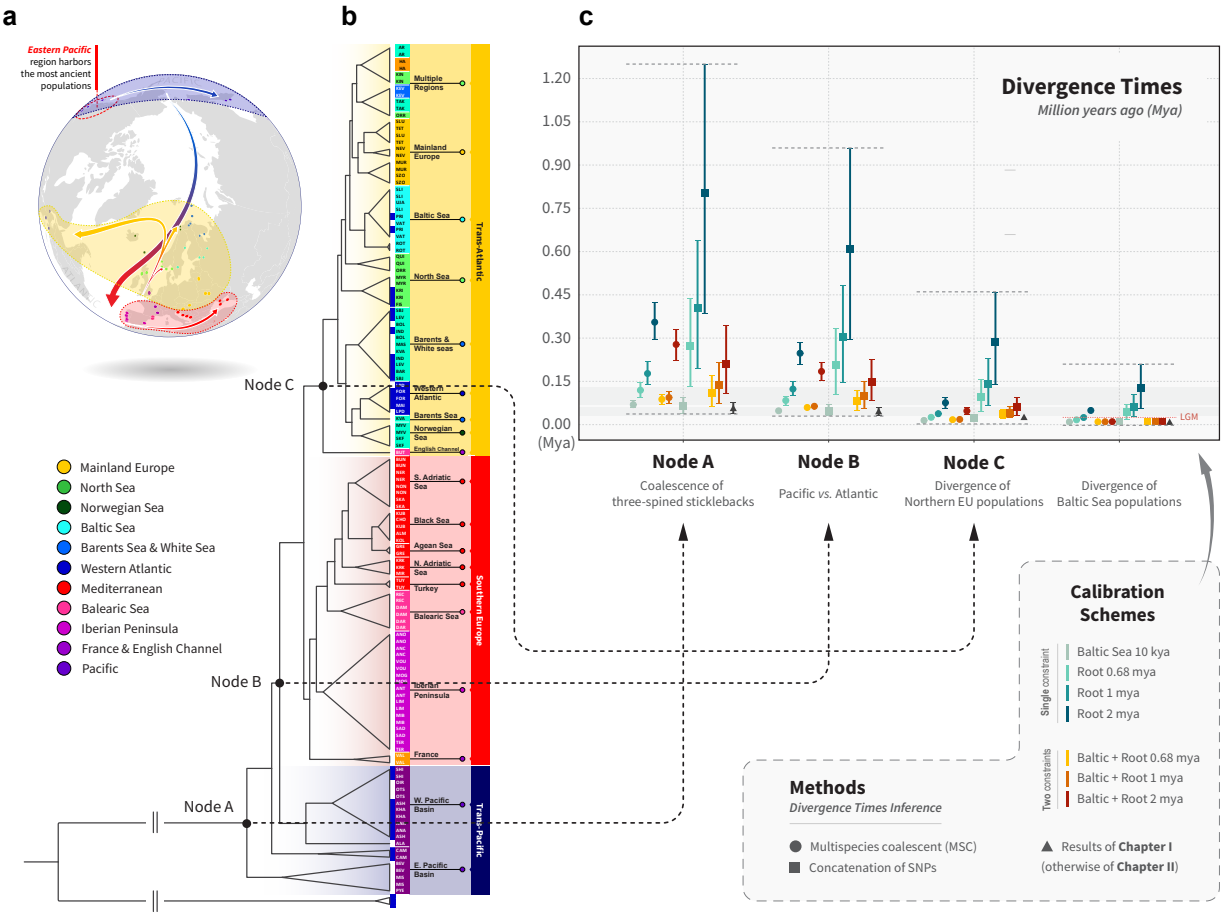


Figure 3 | Worldwide phylogeny of three-spined sticklebacks. (a) Schematic map showing the population localities in different geographic regions. The regions are distinguished by coloured dots (see also Figure 1 in **Chapter I**). The distributions of three major clades are shaded by different colours: Pacific Clade (blue), Southern European Clade (red) and Trans-Atlantic Clade yellow). The arrows indicate hypothesized major colonisation routes. (b) The phylogenetic tree of three-spined sticklebacks with sampling location codes, distribution area and habitat type indicated (**Chapter I**). Marine habitats are labelled with blue rectangles on the left side of the sampling code, the rest of the samples are from freshwater habitats. Closely related populations are collapsed and marked with triangles. (c) Summary of divergence time estimates derived from the multispecies coalescent (MSC)-based and concatenation-based methods (**Chapter II**). The divergence times are estimated based on a subset of 39 individuals, originating from 19 different populations, plus two outgroup samples of *G. nipponicus* used in **Chapter I**. In both methods, seven sets of calibration schemes were applied. The different calibration settings are stated in the Materials and Methods, and marked with different colours. The grey-shaded areas mark periods of Bering Seaway existence. Divergence time estimates are shown as mean ages (dots, squares and triangles) and 95% highest posterior density (HPD) intervals (vertical lines).

zing 1963; Sanz et al. 2015). My results are consistent with the first hypothesis.

Phylogenomics of the nine-spined stickleback and other Pungitius species

The phylogenies of nine-spined sticklebacks and other six *Pungitius* species inferred by the ML-based and MSC-based methods (**Chapter III**) indicated there were two major divergent clades within the genus *Pungitius* (Fig. 4). One of these clades contained the species *P. tymensis*, *P. sinensis* and *P. kaibarae* residing in the Far East (hereafter: sinensis-clade), and the other clade contained the species *P. platygaster*, *P. hellenicus*, *P. laevis* and *P. pungitius* (nine-spined stickleback) inhabiting Eurasia and North America (hereafter: *Pungitius*-clade; Fig. 4). The wider distribution range of the *Pungitius*-clade is explained by the broad distribution of one species, the nine-spined stick-

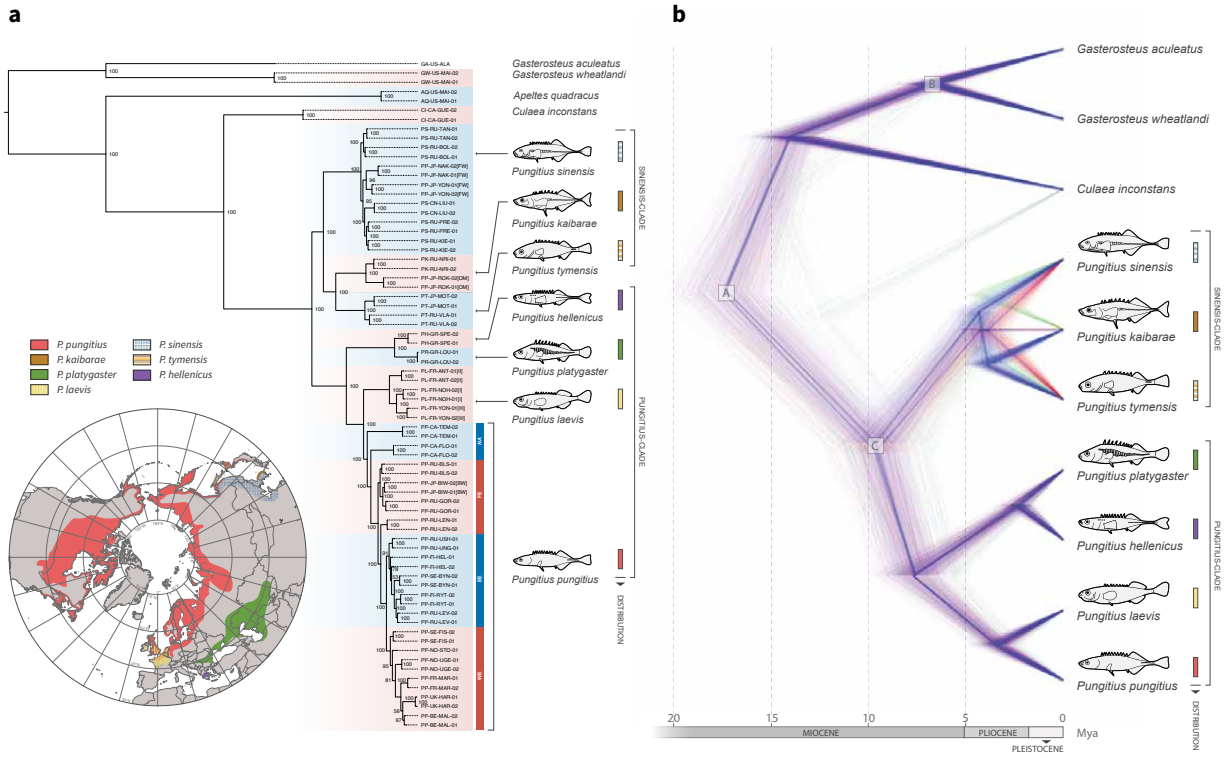


Figure 4 | Phylogenies of *Pungitius* species. (a) Individual level Maximum Likelihood (ML) phylogeny. The numbers beside the nodes indicate posterior probability values. Four lineages of nine-spined sticklebacks (*Pungitius pungitius*) are marked in the phylogeny (NA=North American Lineage, FE=Far East Lineage, EE=Eastern European Lineage, WE=Western European Lineage). The map at the bottom shows the distribution range of each Pungitius species, with the same colour codes as in the phylogenetic trees. The details of the sampling localities are given in Fig. 1 of **Chapter III**. (b) Time-calibrated species-level phylogeny. The phylogeny was estimated based on three calibration points (A, B and C) using the MSC method implemented in the program SNAPP (Bouckaert et al. 2014).

leback, which occurs from North America to Eurasia (Fig. 4). The phylogenies also revealed four major intraspecific lineages of the nine-spined stickleback: North American, Far East, Eastern European and Western European lineages (Fig. 4a).

My data allowed me to infer a broad-scale biogeographical history of the *Pungitius* genus, suggesting that it originated from multiple colonisations from the Far East through the Arctic Sea basin (Fig. 4). The TMRCA of the studied *Pungitius* taxa could be traced back to 7.15–11.61 Mya, in the late Miocene. This is the split time between the sinensis-clade and *Pungitius*-clade, which overlapped with the first opening of the Bering Strait (7.4–4.8 Mya; Marincovich & Gladenkov 1999). The diversification within the sinensis-clade occurred 4.26 Mya in the Pliocene (Fig. 4b; **Chapter III**), through which three species (*P. tymensis*, *P. sinensis* and *P. kaibarae*) evolved in the Far East. However, the diversification history of the *Pungitius*-clade is apparently more complex, as it contained species and populations across the Pacific and Atlantic basins.

Based on the geographical considerations and the inferred phylogenetic affinities, the *Pungitius*-clade evolved two (or more) lineages in the Atlantic 7.15–11.61 Mya (Fig. 4b). One of these lineages colonised towards the Ponto-Caspian area through ice lakes and gave rise to the two southernmost species, *P. platygaster* and *P. hellenicus*. The other lineage split into *P. laevis* and *P. pungitius* about 2.43–4.30 Mya (Fig. 4b; Fig. S3 in **Chapter III**). Among these, *P. laevis* might represent the first wave of trans-Arctic colonisation of Europe.

The diversification of extant nine-spined sticklebacks started from North America about 1.87–3.57 Mya (TMRCA). The common ancestor in the Far East gave rise to the independent colonisations of Europe and North America, and the European nine-spined sticklebacks evolved into two divergent intraspecific lineages that split from each other about 0.43–0.92 Mya (Fig. 4b). The phylogenetic analyses in **Chapter III** also revised the colonisation scenarios for the Western and Eastern Europe lineages (WE and EE) of the nine-spined stickleback. The results reject the hypotheses that the southern European refugia were the source of the EE and WE lineages (Shikano et al. 2010; Teacher et al. 2011). Instead, the results support the hypothesis that Europe experienced two independent colonisations: the WE lineage invaded from north along the Norwegian coast, and the EE lineage colonised from western Russia through Finland along with the retreating Scandinavian ice sheet.

My results further clarified the long-standing conundrum over *Pungitius* taxonomy. The SNP-based phylogeographic patterns and genetic divergence analyses suggested there were at least seven *Pungitius* species, and the originally recognized "freshwater-" and "omono-" type *P. pungitius* actually correspond to P. sinesis and *P. kaibarea*, respectively (**Chapter III**). This updates the affinities suggested by earlier mitochondrial (e.g. Takahashi & Goto 2001; Takata et al. 1987; Wang et al. 2015) and the morphology-based studies (e.g. Takata et al. 1987), verifying that mtDNA and morphology-based systematic inferences in this genus can be misleading (see also: Wang et al. 2015; Takahashi et al. 2016). The results of my genetic analyses also suggest that there may be additional species to be described in this genus: divergent lineages within *P. laevis* and *P. pungitius* might turn out to be different species in studies to come. In fact, one of the three *P. laevis* lineages has recently been described as a new species, *P. vulgaris* (Denys et al. 2018).

Frequent hybridization and introgression events were found among *Pungitius* species on the basis of the incongruence between mitochondrial and nuclear gene trees (mito-nuclear incongruence), as well as on the basis of D-statistic and D_{FOIL} tests (**Chapter III**). First, four cases of mito-nuclear incongruence were detected, three of which occurred in the sinensis-clade and one in the *Pungitius*-clade (**Chapter III**). These mito-nuclear incongruences suggested the occurrence of inter-specific admixture and past mitochondrial capture events (i.e., complete replacements of mitochondria of one species with that from the other species) among four out of the seven studied *Pungitius* species. Such high incidence of mito-nuclear incongruence highlights the limitation of mitochondrial markers in taxonomic inference (e.g., Toews & Brelsford 2012). Moreover, D-statistic and D_{FOIL} tests confirmed admixture in the nuclear genomes of these *Pungitius* species. All nuclear introgression cases involved *P. pungitius*, including admixtures between *P. pungitius* and *P. sinensis*, between *P. pungitius* and *P. kaibarae*, and between *P. pungitius* and *P. laevis* lineage III (**Chapter III**).

Geographically heterogeneous parallel evolution in three-spined sticklebacks

My analyses of genome-wide SNP data from global samples of marine and freshwater populations of *G. aculeatus* revealed 0.208% of the genotyped loci to be associated with marine–freshwater genetic parallelism (Fig. 5e-h; see more identified LD-clusters of this category in the **Chapter IV**). However, 10.3 times more loci (2.149%) were involved in marine–freshwater differentiation exclusively for the Eastern Pa-

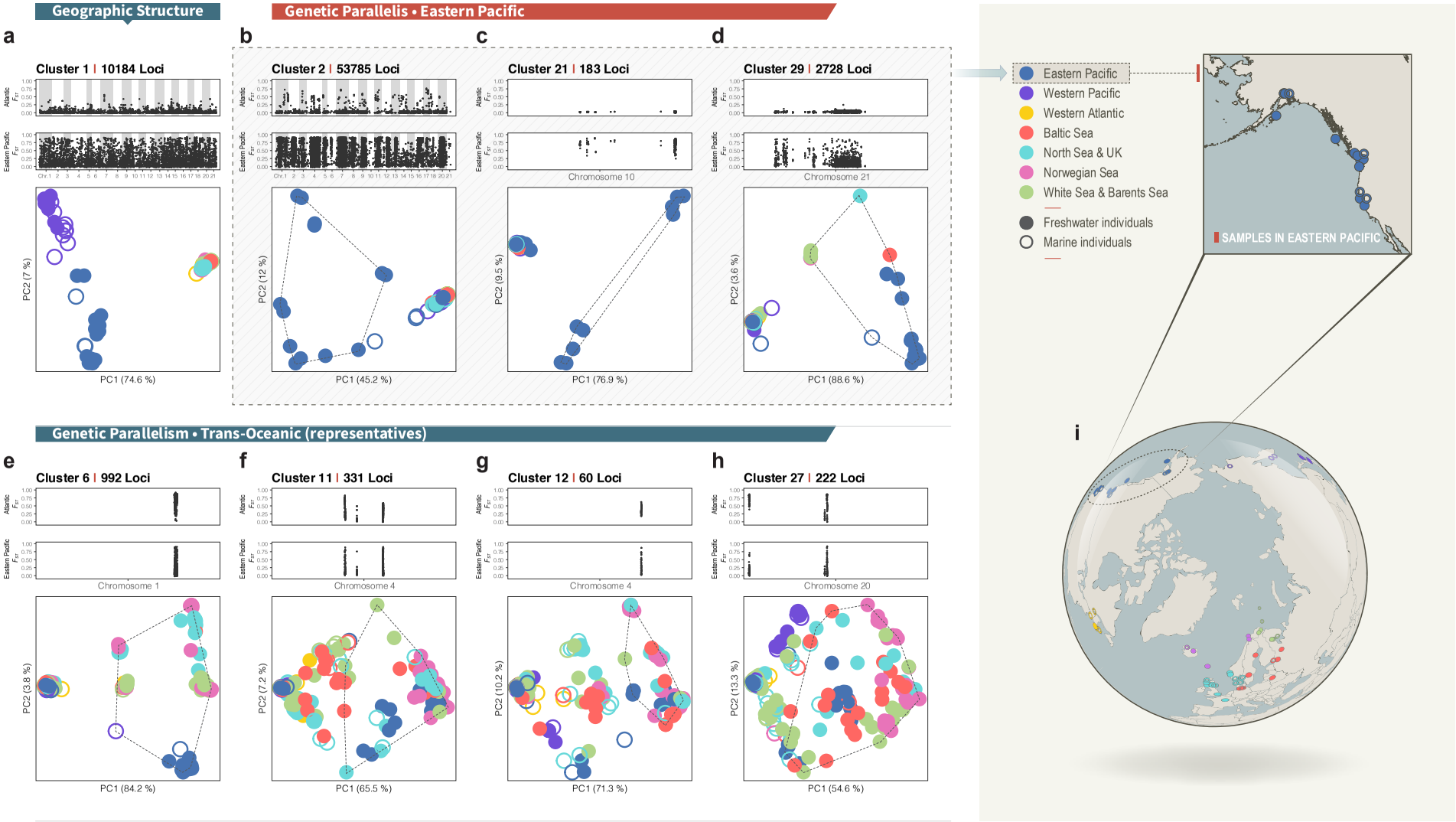


Figure 5 | Marine-freshwater genetic parallelism of three-spined sticklebacks detected by LDna. (a-h) Eight main clusters of loci identified by LDna (LD-clusters). Among them, (a) reflects geographic structure between the Pacific and Atlantic oceans. (b-d) involve marine-freshwater genetic parallelism in the Eastern Pacific. (e-h) represent four LD-clusters responsible for genetic parallelism across the two oceans. In each panel (LD-cluster), the top and middle plots present marine-freshwater differentiation (F_{ST}) between Atlantic and Eastern Pacific samples, respectively. The bottom plot shows the principal component analysis (PCA) based on the LD-cluster loci. The seven different colours in PCAs represent the geographic origin of individuals, which is shown in the map (j). Solid and open circles refer to freshwater and marine ecotypes, respectively. All the identified 29 LD-clusters of worldwide three-spined sticklebacks and their corresponding information are available in **Chapter IV**.

cific populations (Fig. 5b-d). Thus, the results revealed marked heterogeneity in the degree of parallel genetic evolution in three-spined sticklebacks, which was much more pervasive in the Eastern Pacific than anywhere else in the world.

The signatures of genetic parallelism were verified by F_{ST} genome scans. The genomic regions of elevated marine–freshwater differentiation (F_{ST}) were overlapping with the regions identified by LDna analyses that showed signatures of parallel evolution (Fig. 5a-h). These signatures could be seen both in the Eastern Pacific and in the Atlantic regions reflecting global parallelism (top two panels of Fig. 5b-d). My analyses successfully recovered most of the marine–freshwater differentiated genomic regions from the seminal study of Jones et al. (2012), missing only the small genomic regions that had low (or no) sequencing coverage or low ecotype differentiation in my data (**Chapter IV**).

The finding of geographically heterogeneous genetic parallelism of three-spined sticklebacks aligns with the results of recent studies that focus on Atlantic populations (e.g. Ferchaud & Hansen 2016; Liu et al. 2018; Pujolar et al. 2017; Terekhanova et al. 2014; Terekhanova et al. 2019). For example, Ferchaud & Hansen (2016) and Liu et al. (2018) reported little parallelism in local adaptation, most notably in Denmark and Greenland, as compared to that seen in the Eastern Pacific populations. Similarly, Terekhanova et al. (2014, 2019) identified only 21 distinct genomic regions showing consistent marine-freshwater divergence in the White Sea area.

These results are consistent with the LDna analyses in **Chapter IV**, where no LD-cluster showed parallel marine-freshwater divergence exclusively among non-Eastern Pacific populations, and that genomic regions underlying parallel evolution across trans-oceanic regions are a subset of the regions underlying parallel evolution in the Eastern Pacific. Above all, based on the results of this and earlier studies, the patterns of genetic parallelism in three-spined sticklebacks appear to be more locally restricted and involve fewer genomic regions than indicated by studies conducted in the Eastern Pacific region.

Both empirical data and the simulations (see **Chapter IV**) indicate reduced SGV in the evolutionary younger Atlantic marine populations compared to the ancestral Eastern Pacific ma-

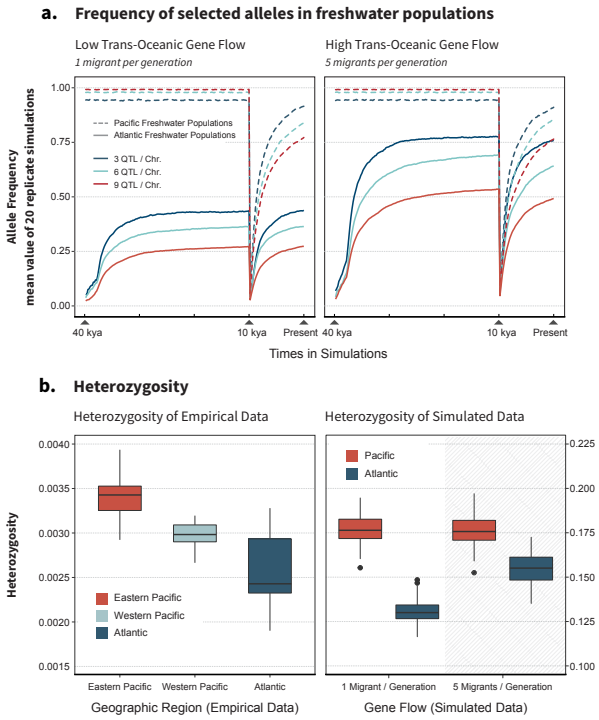


Figure 6 | Ecological genetics of three-spined sticklebacks. (a) Results of genetic simulations showing frequency of freshwater-adapted alleles in the freshwater populations through generations at high and low levels of trans-oceanic gene flow and different QTL-densities (**Chapter IV**). (b) Boxplots of observed heterozygosity in different geographical regions in the empirical and simulated data (empirical data, GLM, $F_{2,64}$=43.05, P<0.001; simulated data: GLM, $F_{1,238}$=509.7, P<0.001; **Chapter IV**). Only trends (rather than absolute values) of heterozygosity should be compared between empirical and simulated data. The simulations were based on the demographic scenario of the "transporter hypothesis" sensu Schluter & Conte (2009; **Box 5**).

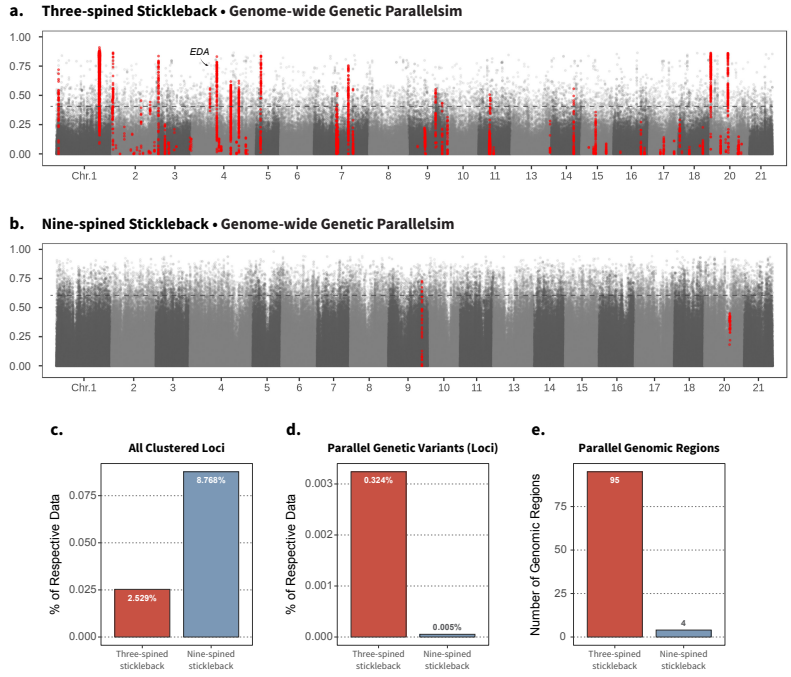


Figure 7 | Quantitative comparison of genetic parallelism between three- and nine-spined sticklebacks. (a, b) Genome-wide allelic marine-freshwater differentiation (F_{ST}) of the Atlantic populations of the two species, respectively. The loci associated with marine-freshwater genetic parallelism identified by LDna are highlighted in red. The dashed line indicates 0.99 quantile of the background F_{ST}. (c) The difference in the amount of loci clustered in LDna between the two species. (d,e) The difference in the amount of genetic parallelism between the two species calculated based on the proportion of genetic variants (d) and genomic regions, defined by a region where at least two LD-cluster loci reside within a window of 50k consecutive sites (e).

rine populations. In my data, individual heterozygosity of Eastern Pacific freshwater-adapted alleles was 29 times lower in the Atlantic as compared to the Eastern Pacific (Fig. 6b; **Chapter IV**). My analyses showed a statistically significant reduction in heterozygosity in the more recently colonised Atlantic region than in the ancestral Eastern Pacific region (generalized linear model [GLM], $F_{2,64} = 43.05$, $P < 0.001$; Fig. 6b). The forward-in-time individual-based population genetic simulations, which followed a demographic scenario mimicking the species' colonisation history (**Box 6a-e**; **Chapter I and II**), predicted the same trend – a reduction in heterozygosity in the Atlantic region compared to that in the Eastern Pacific region (GLM, $F_{1,238} = 509.7$, $P < 0.001$; **Chapter V**).

Schluter and Conte (2009) proposed the 'transporter hypothesis' to account for parallel evolution of freshwater three-spined stickleback populations (**Box 6**). In light of this hypothesis, the lower levels of genetic parallelism in the Atlantic could be explained by the loss of freshwater-adapted alleles due to founder effects and/or selection against the freshwater adapted alleles in the marine environment during the colonisation of the Atlantic Ocean basin (**Box 6**; Fig. 6a).

A possible alternative explanation for the observed high levels of marine–freshwater genetic parallelism in the Eastern Pacific compared to rest of the world is provided by the "secondary contact" hypothesis (Bierne et al. 2013; **Box 6f-g**). Following this hypothesis, the freshwater three-spined sticklebacks from the Eastern Pacific could have originated from an old freshwater population that colonised freshwater habitats by following meltwater from retreating glaciers after the last ice age (**Box 6f**). If the resulting freshwater population then remained isolated from any marine populations (**Box 6f**), the adaptive and neutral genetic differences between marine and freshwater populations could have evolved in allopatry, i.e. in the absence of gene flow. Following a secondary contact between the marine and freshwater populations (**Box 6g**) only after the Atlantic basin was colonised from the Pacific, this scenario could result in extensive genetic parallelism exclusively among freshwater populations in the Eastern Pacific. In support of the "secondary contact" scenario, there is geological

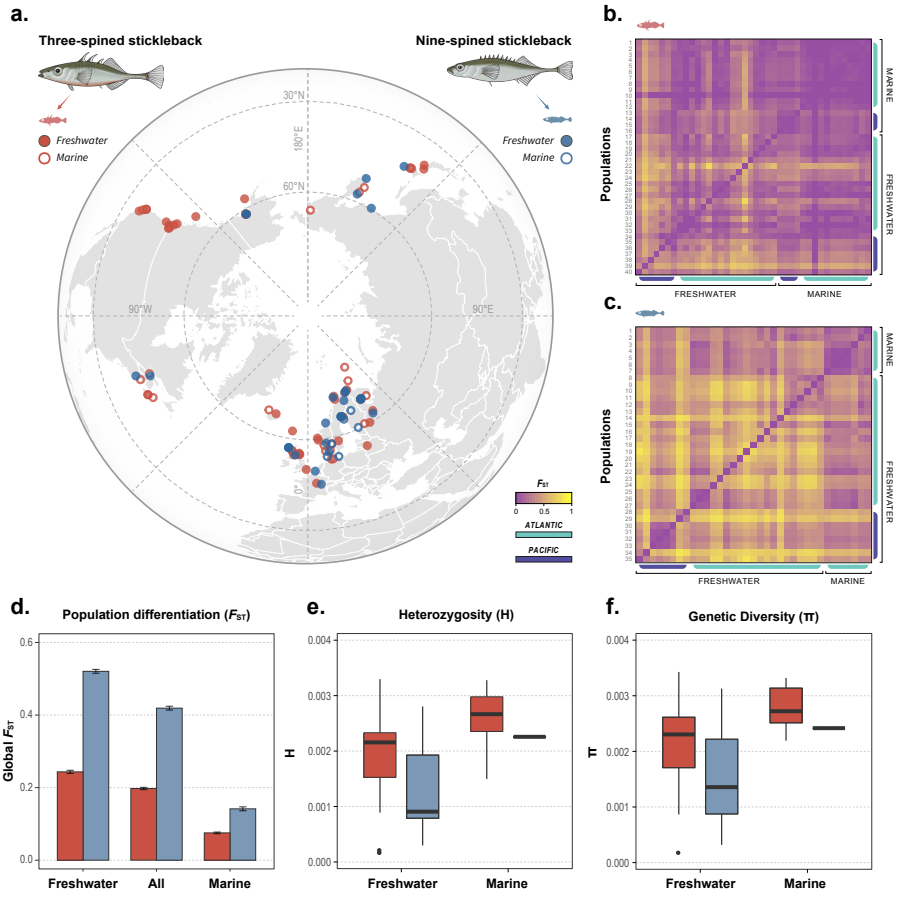


Figure 8 | Genetic variation and population differentiation in three- and nine-spined sticklebacks. (a) Sampling map of the populations used in **Chapter V**. (b, c) Heat map of pairwise population differentiation (F_{ST}) in three- and nine-spined sticklebacks, respectively. (d) Global F_{ST} of the two species (mean ± 95% CI). (e) Boxplot of individual heterozygosity (H) of the two species. (f) Boxplot of nucleotide diversity (π) of the two species. Generalized linear mixed model (GLMM) revealed significant differences in genetic diversity (e, f) between species and between ecotypes (see **Results**).

evidence for existence of ice-lakes in the Eastern Pacific (**Chapter IV**), and my data revealed long-range LD associated with the genetic parallelism Eastern Pacific (Fig. 5b; **Chapter IV**). Consistent with a secondary contact event, Hohenlohe et al. (2012) also found extensive long-range LD among the marine individuals from the Eastern Pacific. However, testing whether the "secondary contact" scenario better explains the patterns of parallel evolution in the Eastern Pacific than the transporter hypothesis is beyond the scope of this thesis.

Contrasting levels of genetic parallelism between three- and nine-spined sticklebacks

My analyses revealed contrasting levels of genetic parallelism between the two ecologically similar and geographically co-distributed stickleback species: there were 95 genomic regions identified to be under parallel evolution in three-spined sticklebacks, whereas only four were identified in the nine-spined stickleback (Fig. 7e; **Chapter V**). Moreover, the proportion of loci (relative to the whole data set) associated with marine–freshwater differentiation across multiple populations was 123.4 times higher in three- than nine- spined sticklebacks (Fig. 7d).

My data suggests that nine-spined sticklebacks have a reduced and more fragmented pool of SGV than the three-spined stickleback (**Chapter V**). First, the levels of genetic differentiation among nine-spined stickleback populations exceeds those of three-spined stickleback populations by a factor of 2.12 (global F_{ST}, Fig. 8d; **Chapter V**). A similar trend was observed in earlier, more geographically restricted studies (DeFaveri et al. 2012; Merilä 2014; but see: Raeymaekers et al. 2017). Second, my analyses revealed 23-fold stronger IBD among marine

nine- than three-spined stickleback populations (Fig. 9; **Chapter V**). Third, three-spined sticklebacks have higher levels of genetic diversity than nine-spined sticklebacks (**Chapter V**). Specifically, the average heterozygosity (H) , nuclear diversity (π) and Watterson's theta (θ) were all significantly higher in the former than the latter species (H: GLMM: $F_{1,258.85}$=91.33, P<0.001; π: GLMM, $F_{1,58.91}$=10.34, P=0.002; θ: GLMM, $F_{1,58.98}$=12.48, P<0.001; Fig. 8e,f). Earlier studies also showed lower levels of genetic diversity in nine-spined than in three-spined stickleback populations (DeFaveri et al. 2012; Merilä 2013). In both species, marine populations had higher genetic diversity than freshwater populations (H: GLMM: $F_{1,257.14}$=25.70, P<0.001; π: GLMM, $F_{1,58.98}$=12.49, P<0.001; θ: GLMM, $F_{1,58.61}$=7.25, P<0.01; Fig. 8e,f). This evidence suggests substantially stronger drift in freshwater than in marine populations, as well as among nine- than three-spined stickleback populations. As such, stronger genetic drift would result in a more heterogeneous pool of SGV available for freshwater adaption in nine- compared to three-spined sticklebacks, hence reducing the probability of parallel evolution.

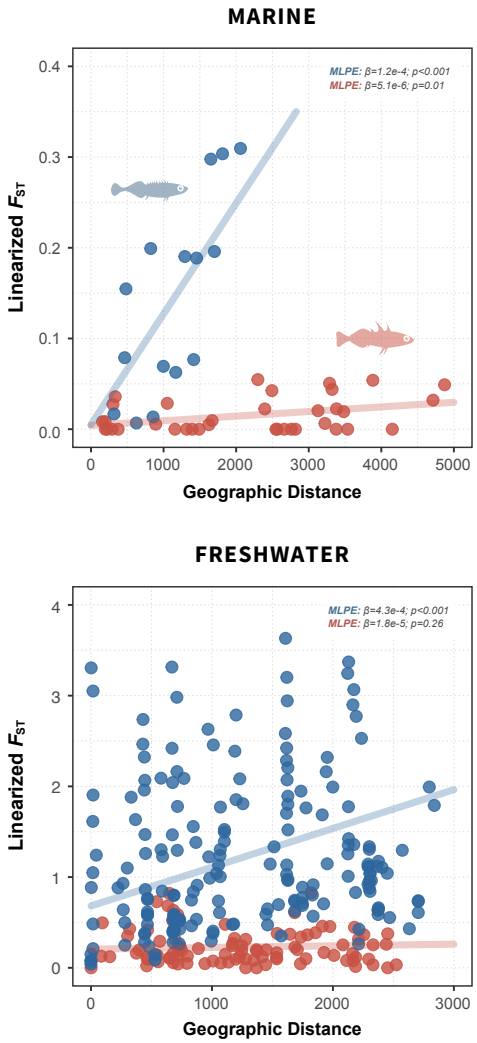


Figure 9 | Isolation by distance (IBD). IBD of the European populations of three- (red) and nine-spined sticklebacks (blue) tested in marine and freshwater habitats, using a maximum-likelihood population-effects (MLPE) model. The results of the regression coefficient (β) suggest stronger IBD in nine- than in three-spined sticklebacks in both marine (23.9 times higher) and freshwater (23.6 times higher) habitats.

Population history also plays an important role in determining the probability of genetic parallelism not only because it affects the contemporary demographic parameters, but also because closely related taxa share more similarities in the environment and genetic characteristics which work to enhance genetic parallelism among them (Conte et al. 2012; Ord & Summers 2015; Rosenblum et al. 2014). Phylogenomic analyses show that the divergence times among lineages of nine-spined sticklebacks are much older than those within three-spined sticklebacks (**Chapters I, II and III**). This is also true in the case of the Atlantic basin populations used to compare levels of genetic parallelism between the two species. Therefore, the older evolutionary history of the nine-spined sticklebacks could have contributed to lower levels of genetic parallelism. In the comparative analyses, I found a significant negative correlation between divergence time and genetic parallelism in nine- but not in three-spined sticklebacks (**Chapter V**), probably because there is much lower variance in divergence time in three-spined sticklebacks. This suggests a significant role for divergence time in affecting the probability of genetic parallelism in taxa with older evolutionary age and/or stronger population structure.

Last but not least, the more heterogeneous pool of SGV in nine- compared to three-spined sticklebacks may also suggest that the former species is subject to less optimal adaptive solutions when adapting to new or changing environmental conditions. This is because the nine-spined stickleback, due to its smaller Ne and restricted gene flow, is less likely than the three-spined stickleback to acquire adaptive variants through gene flow and from new mutations (cf. Barrett and Schluter 2008). Using simulations, Kemppainen et al. (2020) predicted that strong IBD in the sea restricts freshwater populations from reaching their adaptive optima due to limited SGV in the founding marine populations. Therefore, my results predict that with an increasing degree of population subdivision, the probability of gene reuse would decrease, leading to lower potential for local adaptation when colonising novel habitats.

Conclusions and Outlook

A well-defined phylogenetic context is needed to study evolution in replicate populations. The first part of this thesis addressed phylogenetic and demographic hypotheses regarding evolutionary histories and relationships in each species (**Chapters I, II and III**). This knowledge provided the backbone on which the rest of the thesis is built, providing a framework through which hypotheses about parallel evolution can be tested. In the second part of the thesis, I found evidence for strong heterogeneity in the degree of parallel genetic evolution within three-spined sticklebacks, as well as between three- and nine-spined sticklebacks (**Chapters IV and V**). Interestingly, in both cases, heterogeneity in genetic parallelism (repeatability of evolution) could be largely explained by the corresponding distribution of SGV across populations, which is in turn partly explained by the species' demographic and evolutionary history.

The phylogeny of three-spined sticklebacks indicated that their ancestral population resided in the Eastern Pacific region, and the Atlantic Ocean was invaded through the Arctic Ocean and Bering Strait just before LGM. My results in **Chapters I and II** revealed that the species diversification was far more recent than suggested by previous studies and that the current Northern European populations were colonised from a southern refuge post-glacially. The phylogenies of nine-spined stickleback and the other six *Pungitius* species in **Chapter III** indicated that the origin of the genus *Pungitius* resides in the Far East, from where they colonised the rest of their current distribution range via the Arctic Ocean basin. My results also provide evidence for frequent hybridization within the genus *Pungitius*, and clarify taxonomic identities and affinities among

problematic taxa. In general, the results illustrate the power of large genomic data sets in resolving evolutionary and systematics conundrums (**Chapter III**). Although my study established a robust phylogenetic hypothesis of evolutionary relationships in the genus *Pungitius*, future studies should seek to sample more populations of sparsely sampled species (e.g. *P. platygaster*), and include unsampled species described on the basis of their morphology (e.g. *P. bussei* [Eschmeyer et al. 2016] and *P. polyakovi* [Shedko et al. 2005]).

With a focus on parallel evolution in the three-spined stickleback model, I found strong geographic heterogeneity in genetic parallelism in this species – genetic parallelism among Eastern Pacific freshwater populations was 10 times greater than anywhere elsewhere in the world (**Chapter IV**). Such a regional discrepancy was most likely caused by the stochastic loss of freshwater-adapted alleles during historic colonisation, because the genomic regions involved in parallelism in the Atlantic region are a subset of those found in the Eastern Pacific region, and there was a significant reduction of SGV in the Atlantic Ocean compared to the ancestral Eastern Pacific region. These results suggest that the extraordinary levels of parallelism exhibited in the Eastern Pacific – on which most studies are based – are potentially exceptional. As suggested by this phenomenon in three-spined sticklebacks, future work studying parallel evolution in any species might need to consider the generality of the genomic pattern with respect to parallel evolution at broader geographic scales. Moreover, future studies on parallel evolution should carefully assess whether introgression and other demographic events have contributed to unusually high levels of genetic parallelism in the wild. This "secondary contact" scenario can be identified as an alternative explanation for the high level of parallelism in the three-spined sticklebacks in the Eastern Pacific (**Box 6f-g**), but has not been fully explored in this thesis (see results and discussion). To examine this hypothesis, future work would need to incorporate more samples from the Eastern Pacific, and involve coalescent and/or forward-in-time individual-based simulations to rigorously test alternative demographic scenarios.

By comparing population demographic characteristics, evolutionary histories and genetic parallelism between two co-distributed species that are likely to be under similar selective pressure, the results of **Chapter V** indicated that nine-spined sticklebacks exhibit stronger population subdivision, less SGV and much lower levels of parallel evolution than three-spined sticklebacks. These findings suggest that the heterogeneous pool of SGV accessible to nine-spined stickleback demes has resulted in lower levels of genetic parallelism than observed in the three-spined stickleback. This provides evidence that strong population structuring limits parallel evolution. However, my comparative study was restricted to the Atlantic region because of the lack of the marine samples of nine-spined stickleback from the Pacific region. Future studies including samples from Pacific marine populations, as well as incorporation of other stickleback species for a multi-species comparative study, could further clarify the factors influencing the predictability of evolution.

Overall, the results in this thesis indicate that the genetic solutions to similar adaptive challenges are heterogeneous within and across taxa. This likely reflects heterogeneous distribution of SGV within and among species, highlighting the role of evolutionary history and demographic context in affecting adaptive processes in natural populations.

Acknowledgements

My profound gratitude goes first and foremost to my supervisor Professor Juha Merilä. I am most grateful to you for giving me the chance to join EGRU for PhD studies even though population genetics was a new field for me. You generously gave me the time and all the resources for my knowledge to grow in the new field. You were never stingy with your praise and encouragement, no matter how small the progress and the achievement that I made. This really helped me to build confidence and get through the difficulties during the years of my PhD work. Whenever I needed it, you would spare your time from a busy agenda to guide, and support me. It is one of the luckiest things in my life that I became a member of EGRU and was supervised by you.

I would like to express my most heartfelt gratitude to my second supervisor Paolo Momigliano, who spend countless hours and considerable energy in guiding and teaching me. Thank you for raising my knowledge base from layman to specialist. Thank you for your patience when you walked me through the issues step by step and answered my naïve questions. You were always the first one to stand by me when I faced difficulties, and had praise for me when I achieved progress. Whenever I came to you, you would put down your own work and help me. Your optimism and enthusiasm empowered me. Your smiling face is embedded in every pleasant memory in my mind over these years.

I want to convey my whole-hearted appreciation to Petri Kemppainen, my major collaborator and instructor for the work of Chapter IV and the later stage of the my PhD studies. Without you, Chapter IV would not have been greatly improved and in such good shape. During the times with you, I was inspired by your numerous new ideas. Special thanks for your guidance in my coding skills, which improved immeasurably and will benefit me throughout my career.

I would like to thank my Thesis Committee members, Ari Löytynoja and Päivi Onkamo, for your constructive suggestions for my PhD studies. I also want to thank the external examiners of this thesis, Michael Hansen and Walter Salzburger, for your work and positive comments on this thesis.

My sincere appreciation to all EGRU members for your help, support and company during my PhD studies. Special thanks to Miinastiina Issakainen for your lab work and for helping me in sorting out the complicated samplings. To Antoine Fraimout, Emma Vatka and my peer PhD candidates, Sergey Morozov, Mikko Kivikoski and Sunandan Das, many thanks for your wonderful discussions and friendly gatherings on different occasions during my times here. I thank Xueyun Feng for your precious friendship, as we started the PhD studies from China together. Being with you built up my wonderful memories in Finland over these years!

I also want to thank the past EGRU members whom I had chances to work with. I am grateful to Jacquelin De Faveri for your help in the language checking on my several important papers and this thesis. Sometimes you even sacrificed your rest times to help me meet a deadline! Thank Kirsi for your help during my earliest days in Helsinki. I appreciate Cui's friendship and relaxing times during my first two years.

I would like to thank all the countless people at the University of Helsinki who supported me over the years but are not mentioned by name

here. You know who you are but particularly I want to thank Anni Tonteri, the best PhD program advisor in my mind, for your warm-hearted help in my PhD issues. I wish to thank the people from whom I gained friendship when we co-organised the 11th Spring Symposium, attended the workshops in the UK and Italy, and the ESEB 2019. I will remember the wonderful times we have gone through together.

Importantly, I want to express my gratitude to China Scholarship Council, who funded my four-year PhD studies, and the Wuhan University from where I graduated with a Master degree.

Finally, I would like to thank and express my deepest love to my family, in particular my parents. Thank you for your love. I hope I can return it somehow. Thank you for your strong support for my oversea studies. Thank you for your inspiration even when you were experiencing the hardship during the lockdown season in 2020 when I could not be beside you. I love you!

References

Bailey SF, Rodrigue N, Kassen R (2015) The effect of selection environment on the probability of parallel evolution. Molecular Biology and Evolution 32, 1436-1448.

Baker J (1994) Life history variation in female threespine stickleback. In: The evolutionary biology of the threespine stickleback. Oxford Univ. Press.

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. Trends in Ecology & Evolution 23, 38-44.

Bell MA, Foster SA (1994) The evolutionary biology of the threespine stickleback Oxford University Press.

Bierne N, Gagnaire PA, David P (2013) The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Current Zoology 59, 72-86.

Bleidorn C (2017) Phylogenetic Analyses. In: Phylogenomics, pp. 143-172. Springer.

Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE (2018) (Non)Parallel Evolution. Annual Review of Ecology Evolution and Systematics 49, 303-330.

Bouckaert R, Heled J, Kühnert D, et al. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10, e1003537.

Chan YF, Marks ME, Jones FC, et al. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327, 302-305.

Clarke RT, Rothery P, Raybould AF (2002) Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. Journal of Agricultural, Biological, and Environmental Statistics 7, 361.

Colosimo PF, Hosemann KE, Balabhadra S, et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. Science 307, 1928-1933.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. Proceedings of the Royal Society B: Biological Sciences 279, 5039-5047.

Cresko WA, Amores A, Wilson C, et al. (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proceedings of the National Academy of Sciences of the United States of America 101, 6050-6055.

Danecek P, Auton A, Abecasis G, et al. (2011) The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12, 499-510.

DeFaveri J, Shikano T, Ab Ghani NI, Merila J (2012) Contrasting population structures in two sympatric fishes in the Baltic Sea basin. Marine Biology 159, 1659-1672.

DeFaveri J, Shikano T, Merilä J (2014) Geographic variation in age structure and longevity in the nine-spined stickleback (*Pungitius pungitius*). PLoS ONE 9, e102660.

DeFaveri J, Shikano T, Shimada Y, Goto A, Merila J (2011) Global analysis of genes involved in freshwater adaptation in threespine sticklebacks (*Gasterosteus aculeatus*). Evolution 65, 1800-1807.

DeFaveri J, Shikano T, Shimada Y, Merila J (2013) High degree of genetic differentiation in marine three-spined sticklebacks (*Gasterosteus aculeatus*). Molecular Ecology 22, 4811-4828.

DeGiorgio M, Degnan JH (2009) Fast and consistent estimation of species trees using supermatrix rooted triples. Molecular Biology and Evolution 27, 552-569.

Delsuc F, Brinkmann *H,* Philippe *H* (2005) Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6, 361-375.

Denys GPJ, Persat *H,* Dettai A, et al. (2018) Genetic and morphological discrimination of three species of nine-spined stickleback *Pungitius* spp.(Teleostei, Gasterosteidae) in France with the revalidation of *Pungitius* vulgaris (Mauduyt, 1848). Journal of Zoological Systematics and Evolutionary Research 56, 77-101.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Molecular Biology and Evolution 28, 2239-2252.

Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30, 1844-1849.

Edwards SV, Xi Z, Janke A, et al. (2016) Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Molecular Phylogenetics and Evolution 94, 447-462.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends in Ecology & Evolution 26, 298-306.

Eschmeyer WN, Fricke R, Van der Laan R (2016) Catalog of fishes: genera, species, references. Electronic version accessed 19.

Eschmeyer WN, Fricke R, Van der Laan R (2017) Catalog of fishes: genera, species, references.

Ferchaud AL, Hansen MM (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. Molecular Ecology 25, 238-259.

Foster SA (1995) Understanding the evolution of behavior in threespine stickleback: The value of geographic variation. Behaviour 132, 1107-1129.

Fraser BA, Whiting JR (2019) What can be learned by scanning the genome for molecular convergence in wild

populations? Annals of the New York Academy of Sciences 10.1111/nyas.14177.

Futuyma DJ (1998) Evolutionary Biology, Sinauer Associates. Inc. Sunderland, MA.

Gibson G (2005) The synthesis and evolution of a supermodel. Science 307, 1890-1891.

Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit 5.

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5, 184-186.

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology Oxford university press Oxford.

Hendry AP, Peichel CL, Matthews B, Boughman JW, Nosil P (2013) Stickleback research: the now and the next. Evolutionary Ecology Research 15, 111-141.

Herczeg G, Gonda A, Merila J (2009a) Evolution of gigantism in nine-spined sticklebacks. Evolution 63, 3190-3200.

Herczeg G, Gonda A, Merilä J (2009b) Predation mediated population divergence in complex behaviour of nine-spined stickleback (*Pungitius pungitius*). Journal of Evolutionary Biology 22, 544-552.

Herczeg G, Gonda A, Merilä J (2009c) The social cost of shoaling covaries with predation risk in nine-spined stickleback, *Pungitius pungitius*, populations. Animal Behaviour 77, 575-580.

Herczeg G, Turtiainen M, Merila J (2010) Morphological divergence of North-European nine-spined sticklebacks (*Pungitius pungitius*): signatures of parallel evolution. Biological Journal of the Linnean Society 101, 403-416.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. Philosophical Transactions of the Royal Society of London B Biological Sciences 367, 395-408.

Hohenlohe PA, Bassham S, Etter PD, et al. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6, e1000862.

Hohenlohe PA, Magalhaes IS (2019) The population genomics of parallel adaptation: lessons from threespine stickleback. In: Population Genomics: Marine Organisms (eds. Oleksiak MF, Rajora OP), pp. 249-276. Springer International Publishing, Cham.

Jones FC, Grabherr MG, Chan YF, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55-61.

Kapli P, Yang Z, Telford MJ (2020) Phylogenetic tree building in the genomic age. Nature Reviews Genetics 10.1038/s41576-020-0233-0.

Karhunen M, Ovaskainen O, Herczeg G, Merilä J (2014) Bringing habitat information into statistical tests of local adaptation in quantitative traits: A case study of nine-spined sticklebacks. Evolution 68, 559-568.

Kemppainen P, Knight CG, Sarma DK, et al. (2015) Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. Molecular Ecology Resources 15, 1031-1045.

Kemppainen P, Li Z, Rastas P, et al. (2020) Genetic population structure constrains local adaptation and probability of parallel evolution in sticklebacks. bioRxiv https://doi.org/10.1101/2020.01.17.908970.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics 15, 356.

Lambert SM, Reeder TW, Wiens JJ (2015) When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. Molecular Phylogenetics and Evolution 82, 146-155.

Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology 60, 126-137.

Leinonen T, Cano J, Mäkinen *H,* Merilä J (2006) Contrasting patterns of body shape and neutral genetic divergence in marine and lake populations of threespine sticklebacks. Journal of Evolutionary Biology 19, 1803-1812.

Leinonen T, McCairns RJ, Herczeg G, Merila J (2012) Multiple evolutionary pathways to decreased lateral plate coverage in freshwater threespine sticklebacks. Evolution 66, 3866-3875.

Lescak EA, Bassham SL, Catchen J, et al. (2015) Evolution of stickleback in 50 years on earthquake-uplifted islands. Proceedings of the National Academy of Sciences of the United States of America 112, E7204-7212.

Li *H* (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987-2993.

Liu S, Ferchaud AL, Gronkjaer P, Nygaard R, Hansen MM (2018) Genomic parallelism and lack thereof in contrasting systems of three-spined sticklebacks. Molecular Ecology 27, 4725-4743.

Luu K, Bazin E, Blum MG (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. Molecular Ecology Resources 17, 67-77.

Mäkinen HS, Cano JM, Merilä J (2006) Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. Molecular Ecology 15, 1519-1534.

Mäkinen HS, Merilä J (2008) Mitochondrial DNA phylogeography of the three-spined stickleback (*Gasterosteus aculeatus*) in Europe-evidence for multiple glacial refugia. Molecular Phylogenetics and Evolution 46, 167-182.

Mallo D, Posada D (2016) Multilocus inference of species trees and DNA barcoding. Philosophical Transactions of the Royal Society of London B Biological Sciences 371.

Marincovich Jr L, Gladenkov AY (1999) Evidence for an early opening of the Bering Strait. Nature 397, 149.

Marques DA, Jones FC, Di Palma F, Kingsley DM, Reimchen TE (2018) Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. Nature Ecology and Evolution 2, 1128-1138.

McCune AR, Schimenti JC (2012) Using genetic networks and homology to understand the evolution of phenotypic traits. Current Genomics 13, 74-84.

McLennan DA, Mattern MY (2001) The phylogeny of the Gasterosteidae: combining behavioral and morphological data sets. Cladistics 17, 11-27.

Merila J (2014) Lakes and ponds as model systems to study parallel evolution. Journal of Limnology 73, 33-45.

Merilä J (2013) Nine-spined stickleback (*Pungitius pungitius*): an emerging model for evolutionary biology research. Annals of the New York Academy of Sciences 1289, 18-35.

Morales HE, Faria R, Johannesson K, et al. (2019) Genomic architecture of parallel ecological divergence: beyond a single environmental contrast. Science Advances 5, eaav9963.

Morris D (1951) Homosexuality in the Ten-Spined Stickleback (Pygosteus *Pungitius* L.) 1. Behaviour 4, 233-261.

Morris D (1955) The causation of pseudofemale and pseudomale behaviour: a further comment. Behaviour 8, 46-56.

Münzing J (1963) The evolution of variation and distributional patterns in European populations of the three-spined stickleback, *Gasterosteus aculeatus*. Evolution 17, 320-332.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76, 5269-5273.

Nelson TC, Cresko WA (2018) Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evolution Letters 2, 9-21.

Nichols R (2001) Gene trees and species trees are not the same. Trends in Ecology & Evolution 16, 358-364.

Ojaveer E, Pihu E, Saat T (2003) Fishes of Estonia Estonian Academy Publishers.

Ord TJ, Summers TC (2015) Repeated evolution and the impact of evolutionary history on adaptation. BMC Evolutionary Biology 15, 137.

Orr HA (2005) The probability of parallel evolution. Evolution 59, 216-220.

Orti G, Bell MA, Reimchen TE, Meyer A (1994) Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. Evolution 48, 608-622.

Östlund-Nilsson S, Mayer I, Huntingford FA (2006) Biology of the three-spined stickleback CRC Press.

Pease JB, Hahn MW (2015) Detection and polarization of Introgression in a five-taxon phylogeny. Systematic Biology 64, 651-662.

Pembleton LW, Cogan NOI, Forster JW (2013) StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Molecular Ecology Resources 13, 946-952.

Philippe *H,* Delsuc F, Brinkmann *H,* Lartillot N (2005) Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36, 541-562.

Pujolar JM, Ferchaud AL, Bekkevold D, Hansen MM (2017) Non-parallel divergence across freshwater and marine three-spined stickleback *Gasterosteus aculeatus* populations. Journal of Fish Biology 91, 175-194.

Raeymaekers JAM, Chaturvedi A, Hablutzel PI, et al. (2017) Adaptive and non-adaptive divergence in a common landscape. Nature Communications 8, 267.

Ralph P, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 186, 647-668.

Ralph PL, Coop G (2015a) Convergent evolution during local adaptation to patchy landscapes. PLoS Genetics 11, e1005630.

Ralph PL, Coop G (2015b) The role of standing variation in geographic convergent adaptation. American naturalist 186 Suppl 1, S5-23.

Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43, 304-311.

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645-1656.

Rennison DJ, Delmore KE, Samuk K, Owens GL, Miller SE (2020) Shared Patterns of Genome-Wide Differentiation Are More Strongly Predicted by Geography Than by Ecology. American naturalist 195, 192-200.

Rosenblum EB, Parent CE, Brandt EE (2014) The Molecular Basis of Phenotypic Convergence. Annual Review of Ecology, Evolution, and Systematics 45, 203-226.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145, 1219-1228.

Sackton TB, Clark N (2019) Convergent evolution in the genomics era: new insights and directions. Philosophical Transactions of the Royal Society of London B Biological Sciences 374, 20190102.

Sanz N, Araguas RM, Vidal O, Viñas J (2015) Glacial refuges for three-spined stickleback in the Iberian Peninsula: mi-

tochondrial DNA phylogeography. Freshwater Biology 60, 1794-1809.

Schluter D (2000) The ecology of adaptive radiation OUP Oxford.

Schluter D, Conte GL (2009) Genetics and ecological speciation. Proceedings of the National Academy of Sciences of the United States of America 106, 9955-9962.

Schluter D, Marchinko KB, Barrett RDH, Rogers SM (2010) Natural selection and the genetics of adaptation in threespine stickleback. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2479-2486.

Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. Proceedings of the National Academy of Sciences of the United States of America 103, 13753-13758.

Shapiro MD, Marks ME, Peichel CL, et al. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428, 717-723.

Shedko SV (2005) *Pungitius polyakovi* sp. n., a new species of ninespine stickleback (Gasterosteiformes, Gasterosteidae) from south-eastern Sakhalin Island. Flora and fauna of Sakhalin Island 2, 223-233.

Shikano T, Laine VN, Herczeg G, Vilkki J, Merilä J (2013) Genetic architecture of parallel pelvic reduction in ninespine sticklebacks. G3 (Bethesda) 3, 1833-1842.

Shikano T, Shimada Y, Herczeg G, Merilä J (2010) History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations. Molecular Ecology 19, 1147-1161.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Stange M, Sanchez-Villagra MR, Salzburger W, Matschiner M (2018) Bayesian Divergence-Time Estimation with Genome-Wide Single-Nucleotide Polymorphism Data of Sea Catfishes (Ariidae) Supports Miocene Closure of the Panamanian Isthmus. Systematic Biology 67, 681-699.

Stern DL (2013) The genetic causes of convergent evolution. Nature Reviews Genetics 14, 751-764.

Stuart YE, Veen T, Weber JN, et al. (2017) Contrasting effects of environment and genetics generate a continuum of parallel evolution. Nature Ecology and Evolution 1, 158.

Takahashi *H,* Goto A (2001) Evolution of East Asian ninespine sticklebacks as shown by mitochondrial DNA control region sequences. Molecular Phylogenetics and Evolution 21, 135-155.

Takahashi *H,* Moller PR, Shedko SV, et al. (2016) Species phylogeny and diversification process of Northeast Asian *Pungitius* revealed by AFLP and mtDNA markers. Molecular Phylogenetics and Evolution 99, 44-52.

Takata K, Goto A, Yamazaki F (1987) Genetic differences of *Pungitius pungitius* and *P. sinensis* in a small pond of the Omono River System, Japan. Japanese Journal of Ichthyology 34, 384-386.

Teacher AGF, Shikano T, Karjalainen ME, Merilä J (2011) Phylogeography and genetic structuring of European ninespined sticklebacks (*Pungitius pungitius*)—mitochondrial DNA evidence. PLoS ONE 6, e19476.

Telford MJ, Budd GE, Philippe *H* (2015) Phylogenomic insights into animal evolution. Current Biology 25, R876-R887.

Terekhanova NV, Barmintseva AE, Kondrashov AS, Bazykin GA, Mugue NS (2019) Architecture of Parallel Adaptation in Ten Lacustrine Threespine Stickleback Populations from the White Sea Area. Genome Biology and Evolution 11, 2605-2618.

Terekhanova NV, Logacheva MD, Penin AA, et al. (2014) Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus*. PLoS Genetics 10, e1004696.

Thawornwattana Y, Dalquen D, Yang Z (2018) Coalescent analysis of phylogenomic data confidently resolves the species relationships in the Anopheles gambiae species complex. Molecular Biology and Evolution 35, 2512-2527.

Thompson KA, Osmond MM, Schluter D (2019) Parallel genetic evolution and speciation from standing variation. Evolution Letters 3, 129-141.

Toews DP, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology 21, 3907-3930.

Varadharajan S, Rastas P, Loytynoja A, et al. (2019) A High-Quality Assembly of the Nine-Spined Stickleback (*Pungitius pungitius*) Genome. Genome Biology and Evolution 11, 3291-3308.

Wang C, Shikano T, Persat *H,* Merilä J (2015) Mitochondrial phylogeography and cryptic divergence in the stickleback genus*Pungitius*. Journal of Biogeography 42, 2334-2348.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7, 256-276.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38, 1358-1370.

Wootton RJ (1976) Biology of the sticklebacks Academic Press.

Wootton RJ (1984) A functional biology of sticklebacks Springer Science & Business Media.

Xu B, Yang Z (2016) Challenges in species tree estimation under the multispecies coalescent model. Genetics 204, 1353-1368.

Zimmermann T, Mirarab S, Warnow T (2014) BBCA: Improving the scalability of* BEAST using random binning. BMC Genomics 15, S11.